THE ORIGIN AND EVOLUTION OF THE RIBOSOME

BY

AJITH HARISH

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biology
with a concentration in Physiological and Molecular Plant Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Gustavo Caetano-Anollés, Chair and Director of Research
Associate Professor Jay E. Mittenthal
Assistant Professor Matthew Hudson
Professor Lila O. Vodkin

ABSTRACT

The ribosome coordinates one of the most fundamental biological processes, protein biosynthesis. The evolutionary history of the ribosome has intrigued biologists for decades and it's understanding has been a major and elusive challenge.

In this dissertation, the evolution of the ribosome is studied using a novel method that directly embeds structure and function of macromolecules into phylogenetic analyses. Macromolecular structure is more evolutionarily conserved than sequence due to constraints arising from the intricate structure-function relationship. Tracing the evolution of the structural elements of the ribosome, namely ribosomal RNA (rRNA) and ribosomal proteins (r-proteins) provided for the first time to our knowledge, phylogenetic support to hypotheses of ribosome evolution. This study has thus yielded a deep and the most comprehensive insight possible yet into the origins and evolution of the ribosome.

Using phylogenetic methods that reconstruct the evolutionary history of complex RNA and protein ensembles directly from their structure, we find the structures linked to ribosomal processivity that originated in the small subunit (SSU) evolved earlier than the catalytic centre in the large subunit (LSU). Molecular rRNA timelines and phylogenomic analysis of protein structure also show that ancient substructures of the rRNA subunits coevolved with ribosomal proteins at first independently, starting with interactions between the most ancient ribosomal proteins (S12 and S17) and the most ancient small subunit substructure (helix h44) and culminating with the evolutionary integration of the two subunits and most ribosomal proteins to form a modern functional proto-ribosome. These ancient rRNA structures have similarities to *in vitro* evolved RNA ligase and polymerase ribozymes. This indicates that structural elements and functional strategies of a primitive replication mechanism was recruited or co-opted for the process of protein biosynthesis. Functionally important and conserved regions of the ribosome are therefore relics of an ancient ribonucleoprotein world, supporting theories that translation was a functional takeover of a primitive replication apparatus.

To my parents
Harish and Shakuntala
for everything

ACKNOWLEDGEMENTS

Many people have shaped my thinking and helped me learn about "Life", both with regard to life as defined by scientists and life in general. I would like to thank many of them who have helped me at various times before and during my PhD.

Foremost, I would like to express my gratitude to my advisor Gustavo Caetano-Anollés for providing me an opportunity to work with him on an amazing thesis! For constant encouragement, inspiration and patience with me throughout my stay. For financial support through most part of my thesis research.

I thank members of my committee Jay Mittenthal, Matthew Hudson and Lila Vodkin for their patience with my thesis writing till the last few days! Jay, for all the constructive criticisms, numerous discussions and many useful ideas throughout my thesis work. Special thanks to Matt for making it to my defense in spite of the terrible car accident! Lila, for guiding me through all the departmental logistics.

I am very grateful to two other professors, Charles Kurland and Carl Woese, who have inspired and immensely influenced my approach to Science. Chuck, for infecting me with his enthusiasm about ribosomes, for encouraging to think beyond the RNA World and for arranging opportunities to continue working with ribosomes after this thesis. Carl Woese, for inspiring me to think about evolution and helping me realize Science should importantly be pursued not just for useful applications but for a 'greater understanding'.

I would like to thank all my lab mates. Hee Shin for his friendship, Kyung Mo for help with many scripts that made much of my analyses faster and easier, Luda for support and many practical suggestions- scientific and otherwise, Minglei for help with the protein data, Fengjie for his help and support, Suengwoo for Perl scripts and everybody for the numerous critiques and suggestions during lab meetings.

Many thanks to my brother, Rohit Harish, for keeping me afloat financially during the final year of my thesis writing. Also thanks to Amit, Hee Shin and Luda.

I am also grateful to my teachers in India. GR Kantharaj for his inspiring teaching, Manjunath and Savitha for their encouragement, Denny John, Jacob Paul and Bopaiah for introducing me to laboratory experiments.

Thanks to friends at Avesthagen, India for getting me started with my research career. Villoo Patell for giving me an opportuinty, Ganga, Rashmi, Arun, Anthony, Mahesha, Naganand, Naveen, Sanjeev, Gulshan, Renuka, Sulip and Rajesh for training me and for their support and friendship all these years.

Thanks to all my friends in Urbana-Champaign for making it home the last few years. Huzefa for suggesting that I explore an opportunity to work with Gustavo, Amit for driving me home at all times. Govind, Robert, Mallik, Bikash, and Sid for their friendship, their culinary skills, great company, fun-times and all their help and support.

Daniel and Wasi for encouraging and supporting me to get into research when I was less confident to start with.

Sri Devendra and Jagannath uncle, for guidance, support and encouragement.

To my parents! For all their love, sacrifices, for trusting my judgement and letting me embark on a path of my choice and for patience with my grad studies despite not really understanding why it took so long! This thesis is dedicated to my parents.

I gratefully acknowledge support by the National Science foundation (grants MCB-0343126 and MCB-074983607) and the Illinois Campus Research Board.

Table of Contents

Chapter 1
Introduction and Background

1.1 Prologue

The study of evolution is a rather complex endeavor requiring synthesis of knowledge from disparate sub disciplines of biology. Despite tremendous progress, there are many misconceptions associated with the methods that are used to reconstruct evolutionary trees and what they mean. Conventional molecular phylogenetic analyses are based on comparative sequence analysis. The use of molecular structure in evolutionary reconstruction to obtain deep phylogenies is a relatively new method. In addition to the principles of cladistic tree reconstruction, the method incorporates theoretical aspects from thermodynamic studies of molecular structure and thermodynamics of evolution *per se,* which again is rather unconventional. In an attempt to explain why new methods used in this research were developed and how it integrates many theoretical and empirical studies, the introduction that follows has a historical perspective of the concepts and methods used. In addition, it is intended to emphasize the significance of the research presented in this thesis.

1.2 Motivation

The reconstruction of our biological past to understand how we came to be is the most challenging problem of all time [1]. Consequently the evolution of the modern cells is the most important problem in Biology [2]. The Geologic age of Earth is determined to be 4.54 [3] billion years and the life on earth is estimated to have originated 4 billion years ago. Biologists have constantly endeavored to understand the emergence of immense diversity of life and our place on

this planet. Charles Darwin's theory of descent with modification by means of natural selection [4] consolidated many observations into an evolutionary framework used to this day. The advent of molecular phylogenetics and the use of ribosomal RNA (rRNA) as a molecular chronometer extended phylogenetic studies to the microbial world where it was difficult to find distinguishable, observable phenotypes and resulted in the classification of life into a tripartite world [5, 6]. With the recent developments in genomics and bioinformatics, particularly technological advances such as next-generation sequencing, biologists are poised to gain deeper and simultaneously broader understanding about evolution. There has a been renewed focus on the study of evolution with phylogenomics taking center stage in comparative genomics inspiring projects such as Assembling the Tree of Life for all major lineages of life [7].

The choice of rRNA as a chronometer stemmed from efforts towards understanding bacterial evolution led by Carl Woese [8]. rRNAs are ubiquitous molecules in extant cellular organisms. Their functions are universal and highly conserved. They are large molecules that also have relatively less conserved (faster evolving) regions. Thus, both distant and close relationships can be measured using the same chronometer [8]. Furthermore, as part of the ribosome, a highly integrated RNA-protein (RNP) complex coordinating protein synthesis (translation), it is least amenable to horizontal gene transfer and thus portrays true genealogies. In addition to clarifying bacterial phylogenies, rRNA sequence based methods determined a universal phylogenetic tree (UPT) providing a necessary framework to address the problem of evolution of cells (universal ancestors) represented by the root of the UPT. Understanding the evolution of translation is thus essential to understand cellular evolution. It is therefore fundamentally important to understand the evolution of the ribosome.

Despite many advances the root of the tree of life has been intractable due to the inherent limitations with conventional sequence based phylogenies and hence the nature of a universal ancestor is a subject of intense (heated) debates [9]. In conventional phylogenies, outgroups – lineages that fall outside a group of closely related lineages – are used to root a tree of that group. For the UPT however there is no outgroup. Molecular structures are more conserved than sequences and are constrained by thermodynamics [10]. A novel and important phylogenetic method that was recently developed utilizes RNA structure and allows direct reconstruction of

evolutionary histories of molecular diversification [11, 12]. Thermodynamic properties constraining the evolution of molecular structure are invoked to polarize features under examination to produce intrinsically rooted trees and solve the problem of outgroups. In addition several groups have developed phylogenomic methods to build congruent tripartite trees from features describing the occurrence and abundance of protein fold architectures in fully sequenced genomes [12]. In this thesis I have used these two methods to understand the evolutionary history of the ribosome by analyzing its structural components to determine the relative age of each component resulting in a chronology of development of the functional regions.

1.3 The central dogma of molecular biology and the origin of life

Francis Crick in 1958 first proposed the central dogma of molecular biology to explain protein synthesis and the flow of genetic information from DNA to protein sequences involved therein.

> *The Central Dogma*
>
> This states that once 'information' has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the *precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein [13].

The central dogma has often been misunderstood [14] perhaps due to a generalized simplification of the above explanation that implies unidirectional flow of information that is shown below.

DNA > RNA > Protein

Although it was meant to explain gene expression in terms of transcription and translation it influenced thinking about origins of life. If the genetic code harbored in DNA requires protein enzymes for its replication, transcription and translation, but protein synthesis requires information encoded in DNA, how did life originate? This was the classic "chicken or egg" problem. What came first, nucleic acid or protein? [15].

More than ten years after the initial proposal, Crick reiterated to emphasize that the central dogma was formulated to explain transfer of information between sequences of polymers

residue-by-residue, and it was applicable only to extant organisms and not to events that could have occurred during the origins of the genetic code or origins of life [14]. However, the central dogma continues to be misunderstood in the scientific community and also deliberately misconstrued by opponents of the study of evolution, for non-scientific endeavors!

Nonetheless, the origin of translation is certainly a "chicken or egg" problem. Among the three central processes described by the central dogma namely replication, transcription and translation, only translation is mediated by large complexes made up of both RNA and proteins while the former two are carried out by complexes entirely made up of proteins. All three require many additional proteins at various stages of their respective processes. If ribosomes are made up of both RNA and protein, and are required to mediate translation, how did they come to be?

1.4 Phylogenetics, fossils and *in vitro* evolved doppelgängers

1.4.1 A brief history of systematics and phylogenetics

Biodiversity is generally quantified as the number of species and currently the total number has been estimated to lie between 3.6 and 100 million based on various methods, although the consensus is 10 million [16]. Philosophers and scientists alike have long struggled to understand this vast diversity of life. Aristotle is credited to be the first academic biologist as such and to have created a system of biological classification [17]. It was a rudimentary classification of animals alone into Blooded Animals, Non-Blooded animals and Dualizers (sharing more than one property). Although more than two millennia old and now obsolete, it bears resemblance to the modern concepts of "genus" and "species" as Aristotle's system was based on analogous functions of body parts of animals.

In more recent times, exploration of life has been an ongoing effort over the past three centuries. About 1.5 - 1.8 million species have been formally described and cataloged with standard scientific names [16]. Carl Linnaeus, regarded the father of taxonomy, introduced the system of binomial nomenclature and proposed the first nested hierarchical classification scheme

in 1758 in his *Systema Naturae*, broadly classifying life into Animals, Vegetables and Minerals [18]. Although Antony van Leeuwenhoek described microorganisms as early as 1674 in his letters to the *Royal Society*, they were not known widely. Thus living beings were either animals or plants. Taxonomy relied on elaborate description of morphological features to group together similar organisms (assumed to be related) for ease of identification. Their relationships were not evaluated scientifically and thus it was an ordered inventory of organisms. It remained so until 1859, when Charles Darwin and Alfred Russel Wallace's studies showed all life forms were related by common descent and observed diversity was due to natural selection of random variations [19]. In 1866 Ernst Haeckel recognized the distinctive nature of unicellular microorganisms and classified organisms based on their evolutionary relationship to propose a "phylogeny" with three major kingdoms Animals, Plants and Protists [20]. In 1938 Herbert Copeland recognized bacteria to be different and added a fourth kingdom and in 1969 Robert Whittaker proposed a fifth kingdom for fungi [21]. The five-kingdom classification continued to be the accepted standard for more than a century until the advent of molecular phylogenetics. In 1990 Carl Woese proposed the now accepted three Domains of Life [5].

Phylogeny was a term coined by Haeckel in 1866, originally as *Phylogenie* (from Gk. phylon "race" + -geneia "origin," from -genes "born") (www.etymonline.com). Haeckel's tree of life was based on his belief in an anthropocentric hierarchy of the relationships between species [22]. Although a phylogeny, it did not represent the natural genealogy of the organisms represented in the tree. Gradually, methods were developed to quantify the extent of morphological differences, which was believed to reflect evolutionary distance. Two fundamental methods have come to be generally used to express morphological (and later molecular) data as phylogenetic trees, phenetics and cladistics [23]. Both methods attempt to realize the same end result but in very different ways. Phenetics, also known as "numerical taxonomy", is based on a principle of comparison of multiple anatomical features to infer taxonomic relatedness by estimating quantifiable similarities and differences. The method by itself does not deduce phylogenetic relationship but is assigned by the pheneticist and thus can be subjective. It calculates an overall similarity and generates a phenogram (tree), which shows patterns of similarities. Thus a phenogram simply represents the distance of the individual species from the root (overall similarity). In other words a phenogram only shows the

evolutionary distance and does not represent the actual evolutionary path that resulted in the observation. Based on the phenogram, a pheneticist can then determine the branching order to fit the observations. Thus more morphologically similar species will be classified into the same group whether or not they are evolutionarily related. In short, classification by phenetics might not always reflect phylogeny.

Cladistics, unlike phenetics, is specifically intended to reconstruct evolutionary histories [23]. Willi Hennig in 1950 proposed a method to classify organisms based on their shared-derived morphological characters for reconstructing trees that reflect lineal decent. A shared-derived character (synapomorphy) is a feature shared by two closely related groups but not by a distant common ancestor. Features shared with the distant common ancestor are termed shared-primitive (symplesiomorphy). Such features are phylogenetically uninformative. Groups having common shared-derived characters are identified as sister groups and placed together in a clade. Thus a cladogram, unlike a phenogram explicitly shows a phylogenetic reconstruction. Cladistic methods are hence objective and do not assume relationships *a priori* and avoid subjective judgments based on similarities. This makes cladistics amenable to hypothesis testing by rigorous statistical methods and to Popperian philosophy of falsifiability. Thus cladistic methods have superseded phenetic methods and have been the choice for the last five decades. In principle, cladistics can be used to classify anything, including inanimate objects [23]. Beyond biology, in historical linguistics they have found applications in reconstructing chronologies [24] and in behavioral sciences [25].

1.4.2 Fossils, radioactive clocks and dating the history of life

The fossil record, among other things, has provided strong irrefutable evidence for evolution. Although rare and embedded in vast ranges of breadths and depths of the Earths' crust, paleontologists have unearthed an astonishing collection of fossils. The fossil record provides a panoramic view of the types of organisms that existed during different ages of the Earth and a link to the past. The earliest record of recognition that fossils were relics of ancient life is associated with Nicholas Steno who in 1766 [26] explained such possibilities based on similarities between shark teeth and the rocks commonly found at that time. One of the most

important developments was the finding of the Neanderthal skull by Thomas Huxley and the development of a theory of evolution of Hominids. Over time, paleontologists assimilated information about extinct organisms and began to use cladistic methods to explain their findings [26].

Since cladistics is explicitly evolutionary, in addition to explaining observations, it is possible to reconstruct ancestors and predict the direction of change of characters defining them. The resulting models not only depict a static record of changes, but also, importantly, the dynamic evolutionary processes that could have driven those changes. Thus cladistic methods have brought about an amalgamation of paleontology, evolutionary synthesis and systematics into modern *phylogenetics*.

At about the time paleontologists and taxonomists embraced evolutionary theory and developed reliable methods, geologists used breakthroughs in radiometric dating technologies and described the geologic history of the Earth with a definitive calendar. In 1956 Clair Patterson determined the age of the earth using the highly accurate Pb^{207}/Pb^{206} isotopic dating to be $4.55 \pm 0.07 \times 10^9$ years [27]. With this major development it was now possible to date fossils and build a chronology of events that have occurred during the history of life on Earth. The oldest fossils discovered yet, stromatolites are dated to be ~ 3.5 billion years old. Based on this, life on Earth is estimated to have originated 3.8-4.0 billion years ago [28]. Fossils, radioactive clocks and constantly updated historical almanacs make it evident that life on Earth has changed since its inception and is almost as old as the planet.

Despite evolutionary study developing leaps and bounds since it inception, it was not possible to trace evolutionary history all the way back to the most recent common ancestor until mid 1970s [5]. Prior to that evolutionary inferences were based on classical phenotypes, predominantly morphology, which were features observable by the naked eye. The focus of such studies were limited to metazoa and metaphyta [5]. Accordingly, studies of fossil evidence were predominantly of skeletal remains or macroscopic fossils, which date back at most to 0.5-1.0 billion years. Therefore, hypothesis-based reconstructions of evolutionary history could describe only the last quarter of history of life on Earth.

1.4.3 Molecules as fossils, molecular clocks and origins of life

In their classic paper of 1965, Emile Zuckerkandl and Linus Pauling laid the foundations for the study of molecular evolution wherein they justified why biological macromolecules among all natural constituents of life retain the maximum possible evolutionary history in their sequences and structures [29]. Phylogenies derived from molecules alone are therefore the most comprehensive and informative. In an earlier study of haemoglobin, they discovered that amino acid sequence variation in related species approximately reflected the time of divergence of those species. For a given set of homologous proteins, random genetic mutations accumulate at a relatively constant rate and thus the number of differences between them increases with time. Therefore, the number of mutations can be used to estimate time. These ideas have been the basis for the molecular clock hypothesis. While many models have since been proposed to estimate divergence times, there has been no consensus and is a highly contentious issue [30-32].

Although molecular chronometers have not yet been calibrated to even approximately match radioactive clocks, most importantly they have made it possible to trace evolutionary history all the way back to ~ 4.0 billion years and deduce the nature of a universal ancestor. Like fossil records have already shown, molecular phylogenetics has also shown that our biosphere was and is dominated by unicellular organisms, which greatly outnumber multicellular organisms both in terms of biodiversity and biomass [33, 34]. During the past decade, the reconstruction of life by means of single/multiple gene/protein sequence based phylogenetics has gradually paved way for whole genome sequence-based phylogenomics. Nonetheless, as with most other scientific methods, sequence based methods are not without limitations. Some of them are highlighted below and a solution to the problem is discussed.

1.4.4 Molecular structures, common ancestors and the root of the universal tree

Tree reconstruction methods generally do not provide for intrinsic means to root the trees they produce [23, 35]. Unrooted trees only describe the evolutionary lineages but not their trajectories. Rooting is usually an *a posteriori* process and is dependent on many assumptions, which convert unrooted trees into directed trees. Since evolution has proceeded in one direction

with respect to time, the different states of a character follow an order. This evolutionary order is determined by the polarity (direction) of character-state change. That is, to root a tree, it has to be determined which state of a character is pleisiomorphic (ancestral) and which is apomorphic (derived). Determining this order is not simple and often involves many assumptions. The most common method used is the outgroup method. For a given set of closely related taxa, an outgroup is relatively distantly related. It is (reasonably) assumed that the synapomorphic character states in the outgroup are ancestral. Therefore the ingroup relationships are determined with the outgroup character states as "reference". Thus choosing an outgroup requires knowledge of the ingroup taxa or assumptions amounting to it. Choosing an appropriate outgroup can be difficult. Fossils alleviate the problem to an extent, but are not sufficiently available. There are no good outgroups in the case of Tree of Life. Apart from outgroups, molecular clocks are used to root phylogenetic trees and many other methods as well [30, 36]. Since the molecular clock hypothesis assumes that the rate of evolution is constant, the root of the tree is the midpoint of the longest path, which represents the distance between the two most distant taxa in the tree. However, the universality of the molecular clock hypothesis is not always reliable [37].

Although there is no suitable outgroup for the tree of life, a paralogous gene deduced to originate from an ancient gene duplication (Elongation Factor-Tu) has been used to root the universal tree [38] (Fig. 1.1a). The root of the tree is the hypothetical last universal cellular ancestor (LUCA). Since the first proposal in 1989, many genes have been used. In all cases, this method places the root in the Bacteria making Archaea and Eukarya sister groups. This rooting has been well accepted for most part since it explains the assumptions (prejudice) that cellular life evolved from simpler cells [9]. However, rooting based on gene duplications has been found to be unreliable for many reasons including unequal rates of sequence evolution, mutational saturation, and long branch attraction artifacts [39]. In addition, the discovery that horizontal gene transfer is pervasive among prokaryotes and new broader pictures from comparative genomics, which showed that each of the genomes of the three superkingdoms were mosaics of two others' protein coding gene repertoires have led to alternative models of the tree of life (Figure. 1.1b-c). However, Forterre and Philippe have noted that in all the models, life evolved from simple prokaryotes to complex eukaryotes and that such proposals are due to

overconfidence in molecular phylogenies and prejudices about the nature of primitive cells [9]. They also point out that in molecular phylogenies the number of positions in protein or nucleotide sequence alignments that are homologous and contain real ancient phylogenetic signal are very limited. This not only makes it extremely difficult to determine stable characters with enough phylogenetic signal in all groups studied but also makes polarization unreliable either with outgroups or molecular clocks (mutational saturation).

Although it is logical to agree that the evolution from the origin of life to LUCA must have been "simple to complex", both complexification and simplification have occurred during evolution. A model for "complex to simple" cellular evolution has been proposed where eukaryotic-like LUCA evolved by simplification to present day prokaryotes (Figure. 1.1d). This hypothesis is supported by comparison of the components of information processing proteins (translation factors, DNA/RNA polymerases and associated factors). In addition, patterns of orthologous gene replacements between Archaea and Bacteria, and reductive evolution in mitochondria and chloroplasts agree with the model.

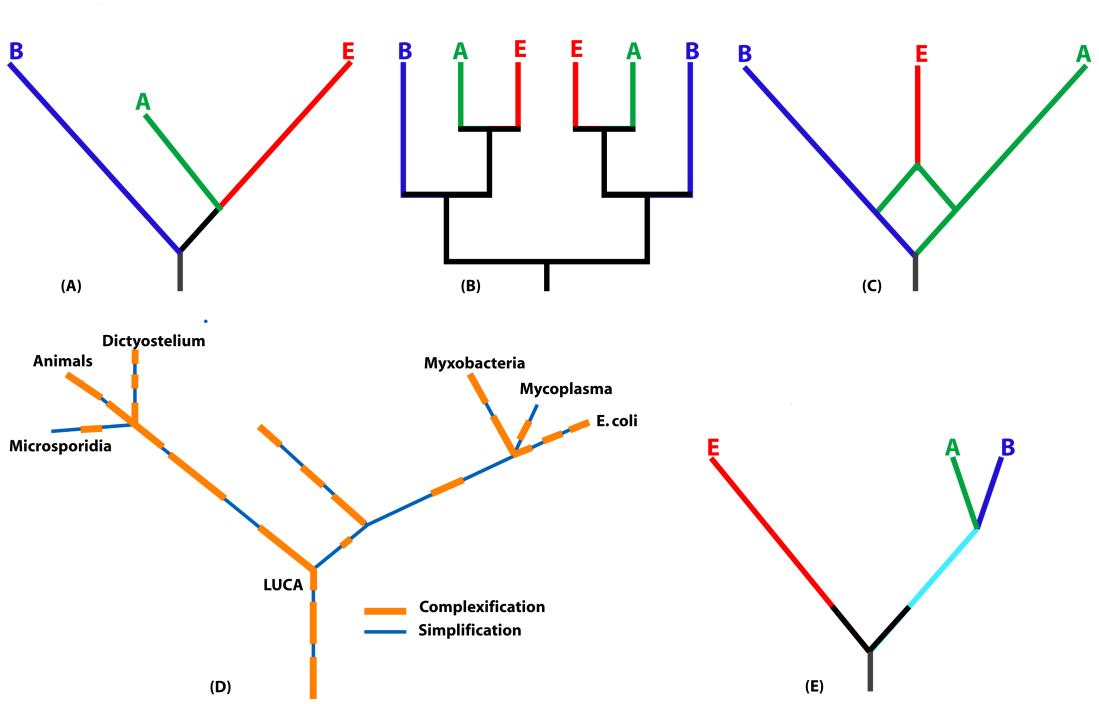


Figure 1.1: Different rooting strategies and the resulting hypotheses. (A) rRNA sequence tree rooted with duplicated gene. (B) Trees of paralogous proteins used to root each other (C) Hypothesis based on gene content. (D) Hypothesis of reductive evolution. (E) rRNA structure based, intrinsically rooted tree. [9]. A representation of an intrinsically rooted tree derived from RNA structural characters agrees with reductive evolution [11]

Zuckerkandl and Pauling in their original proposal recognized that in addition to sequences, 3D structures of proteins could aid tracing the evolutionary history of molecules. Molecular structure is far more conserved than primary sequence, since structures define functions that are ultimately selected during evolution [10]. Thus information in structure unaffected by sequence mutations persists much longer than in sequence. Consequently, phylogenies derived from structural characters are expected to aid the reconstruction of deeper evolutionary histories compared to sequences. Recently in 2002, a novel and important method was developed that reconstructs phylogenies from features that define RNA structure [11]. This method "embeds molecular structure and function directly" into phylogenetic analysis [40] and was awarded the 'Zuckerkandl Prize'. Structural features are treated as ordered multistate characters and invoking an evolutionary tendency towards molecular order polarizes character state transformation. This tendency is well supported by the thermodynamic theory of evolution and principles of statistical mechanics. Phylogenies generated by this method were not only congruent to sequence based methods in realizing the three-domain phylogeny, but also, importantly, are intrinsically rooted. In addition to exploring the intractable root of the tree of life [11] the method has been used to trace the evolution of RNA structure in the ribosome and obtain a rare insight into early evolution of protein synthesis [41].

1.5 Self-Organization, thermodynamics and life as an emergent phenomenon

The study of evolution progressed from Darwin's *Origin of Species* with the broad organismal perspectives of the mid-late 1800s to the 'modern synthesis', where Darwinism and Mendelism were brought together. This probably started with Theodosius Dobzhansky in his *Genetics and Origin of Species* (early-mid 1900s). However, evolution remained conceptually isolated from the physicochemical principles that underlie natural processes [42]. Erwin Schrödinger's *What is Life?* was an attempt to explain the physical (and chemical) basis of life with fundamental principles of thermodynamics [43]. Schrödinger identified that living systems are highly ordered internally and far from equilibrium, an apparent paradox to the Second Law of Thermodynamics. The Second Law of Thermodynamics states that in an isolated, closed system not in equilibrium, entropy tends to increase over time and approaches a state of equilibrium with "maximum entropy". Schrödinger explanation of this paradox was that living systems avoid rapid decay

into an inert state of equilibrium by feeding on 'negative entropy' and phrased it as *order from disorder*. In other words, living systems draw energy from their surroundings and channel it within to decrease entropy and produce internal order.

Many theories have further developed Schrodinger's concepts. Thermodynamics concerns the study of transformation of energy [42, 44]. Free energy is a quantity equivalent to the capacity of a system to do work and entropy is a quantity equivalent to energy unavailable for conversion into work. A biological system exists in larger encompassing system of energy and matter fluxes (due to established gradients). It is basically an energy transformer, where radiant energy is concentrated, and gradually degraded and dissipated (as heat). In the process new structures for efficient dissipation and storage emerge. In the case of the biosphere the free energy gradient from solar energy source and the thermal sink of space sustains the pattern of dissipation. The processes that relieve this gradient by transforming energy to heat are irreversible and thus over time the system evolves through states in increasing complexity and structure (order). Thus, in general, evolution is driven by a unidirectional flow of energy to heat. This directional flow is termed as "thermodynamic arrow". The thermodynamic theory of evolution posits that this thermodynamic arrow drives the overall evolution of the biosphere, from its prebiotic origins through the emergence of life and through the progressive, phylogenetic diversity. Proponents of the theory see *Life as a manifestation of the Second Law of Thermodynamics* [45]. In fact the theoretical framework has been used to even explain cosmological evolution [46].

The theory of self-organization has been tested with RNA secondary structures to assess whether evolution of phenotypes is driven by external constraints or internal constraints [47]. RNA molecules whose functions are structure dependent generally possess well ordered structures that are both thermodynamically stable and uniquely folded. Since RNA folding is understood and can be reliably predicted they are good models to estimate the contributions of selection and self-organization in their secondary structures. In the study, the stability and uniqueness of RNA secondary structures were quantitatively defined by three parameters: (i) base pairing propensity (P), (ii) mean length of helical stems (S), and (iii) uniqueness of the folded structures (Q). P and S measure the stability of a given structures, Q is the probability

that a given sequence can exist in alternate suboptimal conformations close to the most stable conformation. Hence, the smaller the value of *Q*, the more unique is a structure. Figure 1.2 shows the *Q*, *P* and *S* values for a natural RNase P structure and that of those obtained by randomized sequences of the same. Remarkably structures from random sequences had *Q*, *P* and *S* measures similar to the natural RNase P structure. Although evolved conformations were more ordered, the analysis showed that most of the conformational order in evolved sequences is due to intrinsic properties of RNA folding. In summary, a large number of RNA sequences can fold into a much smaller number of closely related structures. A subset of those sequences selected for specific adaptations further evolves, resulting in a very small set of evolved sequences and structures.

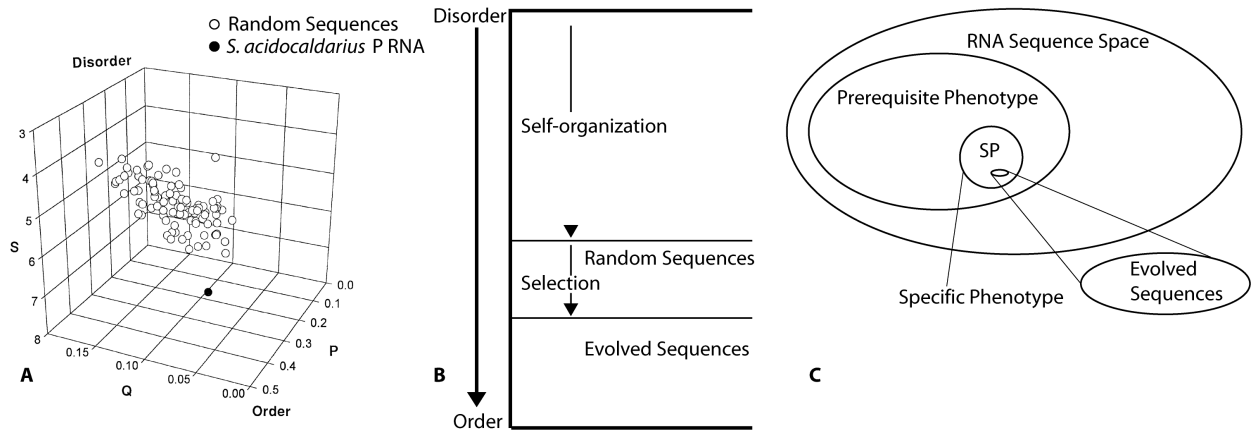


Figure 1.2: Contributions of self-organization and selection in RNA evolved secondary structures. (A) Q, P and S values for *S. acidocalsarius* P RNA and random RNAs (B) A representation of the intrinsic (self-organization) and extrinsic (selection) source of conformational order in evolved RNAs, it shows majority of the conformational order is due to self-organization. (C) Relation between random and evolved sequences. A set well ordered structures (prerequisite phenotype) consist a subset of sequences with specific adaptations within which are a smaller subset of that are selected during evolution [47]

1.6 Gene resurrection, directed evolution and doppelgängers

Most of our understanding of evolutionary processes and mechanisms are from phylogenetic reconstruction and fossil evidence. However, fossil evidence is generally limited and there are few molecular fossils. Consequently, there are many gaps and missing links. More importantly the mechanisms underlying these processes are based on statistical associations of inferred extant

mechanisms in sequence data. However, statistical inferences may not be always reliable, more so when alternative explanations are feasible [48]. The inferences remain unconfirmed hypotheses without any empirical evidence. One way to test the hypotheses is to resurrect gene sequences of ancient proteins/RNA functional molecules based on the inferences from phylogenetic methods. Resurrected gene sequences can be expressed in vitro and characterized to gain insight into primordial adaptations and evolutionary constraints operating on them.

Yet another way is by directed evolution of random sequences from large libraries that selected for specific functions [49]. After repeated rounds of evolution and selection RNA/protein molecules that perform desired functions have been obtained that also provide a window into primordial molecules that likely existed. Importantly, they are proof that biological functions can evolve from random sequences. In addition, it is possible to create the missing 'fossil record' with in vitro evolved catalysts substituting as doppelgängers for extinct molecules. [A doppelganger is ghostly double of a living being or the location of an object in two places (in our case, past and present) at the same time.]

Phylogenetic reconstruction and fossil records constitute a 'top-down' approach while the gene resurrection and directed evolution methods constitute a 'bottom-up' approach to build the UPT with the hope of painting a more comprehensive picture of the evolution of life on earth.

1.7 Overview of the thesis

This thesis consists of three research chapters and a conclusion with a synthesis of the results. Phylogenetic methods that were developed particularly to describe and extract deep phylogenetic signals imprinted in features of molecular structure are used to reconstruct the evolutionary history of the ribosome. These methods axiomatically incorporate a large body of theoretical and empirical evidence to invoke polarization of character state transformations reflecting a tendency towards increased molecular order and stability.

In chapter 2, using information in rRNA secondary structure a universal tree of rRNA structural components was reconstructed to determine the relative ages of the different regions of

the rRNA. The estimated relative age was then used to build a chronology of development of different functions of the ribosome. Unlike previous studies that could only partially analyze the evolutionary history of rRNA, deconstructing the rRNA structure into its structural components made it possible for the first time to comprehensively analyze the evolution of the complete rRNA structure. In addition to confirming previously known hypotheses based on sequence and biochemical data, a new view of ribosomal history was therefore possible.

In chapter 3, the relative ages of conserved ribosomal proteins (r-proteins) were determined from data obtained from a universal tree of protein architecture at the fold superfamily level and used to determine if rRNA and r-proteins co-evolved and if so how primitive was this cooperativity. The coevolution of r-proteins and rRNA was found to be surprisingly older than previously thought and provided new insights into the nature of a primitive ribosome.

Due to the high degree of cooperativity between a multi-component ribosome and its requirement of external factors, the evolution of the ribosome is one of the toughest problems. Some hypotheses have been proposed that the only way such a complex function could have evolved was from a previously related function [50]. In chapter 4, using the results from the previous chapters, new structure comparison tools were employed in an attempt to bridge the missing gap between the top-down and bottom-up evolutionary approaches. This analysis corroborated models that proposed that protein synthesis was a functional take over of a fundamental ancient function.

Finally, Chapter 5 is a synthesis of the results of this research and its implications for our understanding of the origins and evolution life. Results are placed within a broader perspective.

Note: In this thesis, *superkingdom* is used in place of *domain* of Life to avoid confusion with molecular structural domains such as a proteins domain or a RNA domain.

Chapter 2
Origin and Evolution of ribosomal RNA

2.1 Introduction

The ribosome is responsible for the synthesis of nearly every molecule in the cell, either directly or by enzymes made by it [51]. Translation is the final phase of genetic information transfer from DNA to protein as described by the 'central dogma of molecular biology' [13]. Before this, information in DNA sequence is converted to RNA by transcription, which specifies the amino acid sequence of the protein as the triplet genetic code. Translation involves two major aspects, the decoding of the information and catalyzing the peptide bond synthesis. The ribosome catalyzes this process efficiently by increasing the rate of peptide bond formation by $> 10^5$ fold while maintain a high level of accuracy of $> 10^6$ [52]. Until 2000, much of what was known about ribosome functions was by biochemical experiments. The availability of high resolution crystal structures starting in 2000 has dramatically altered the understanding of ribosomal functions and design of better experiments [52]. This in turn has provided a great opportunity to decipher the evolutionary history imprinted in the structure of the ribosome and advance our understanding of the origins and evolution of translation.

2.2 Overview of the structure of the ribosome and its function in translation

Ribosomes are large RNA-protein (RNP) molecular machines that catalyze protein synthesis by translating the genetic information in messenger RNA (mRNA). Transfer RNAs (tRNA) charged with amino acids are substrates for the ribosome. Cytoplasmic ribosomes in all

organisms are made of two subunits, a small subunit (SSU, 30S/40S) and a large subunit (LSU, 50S/60S), each

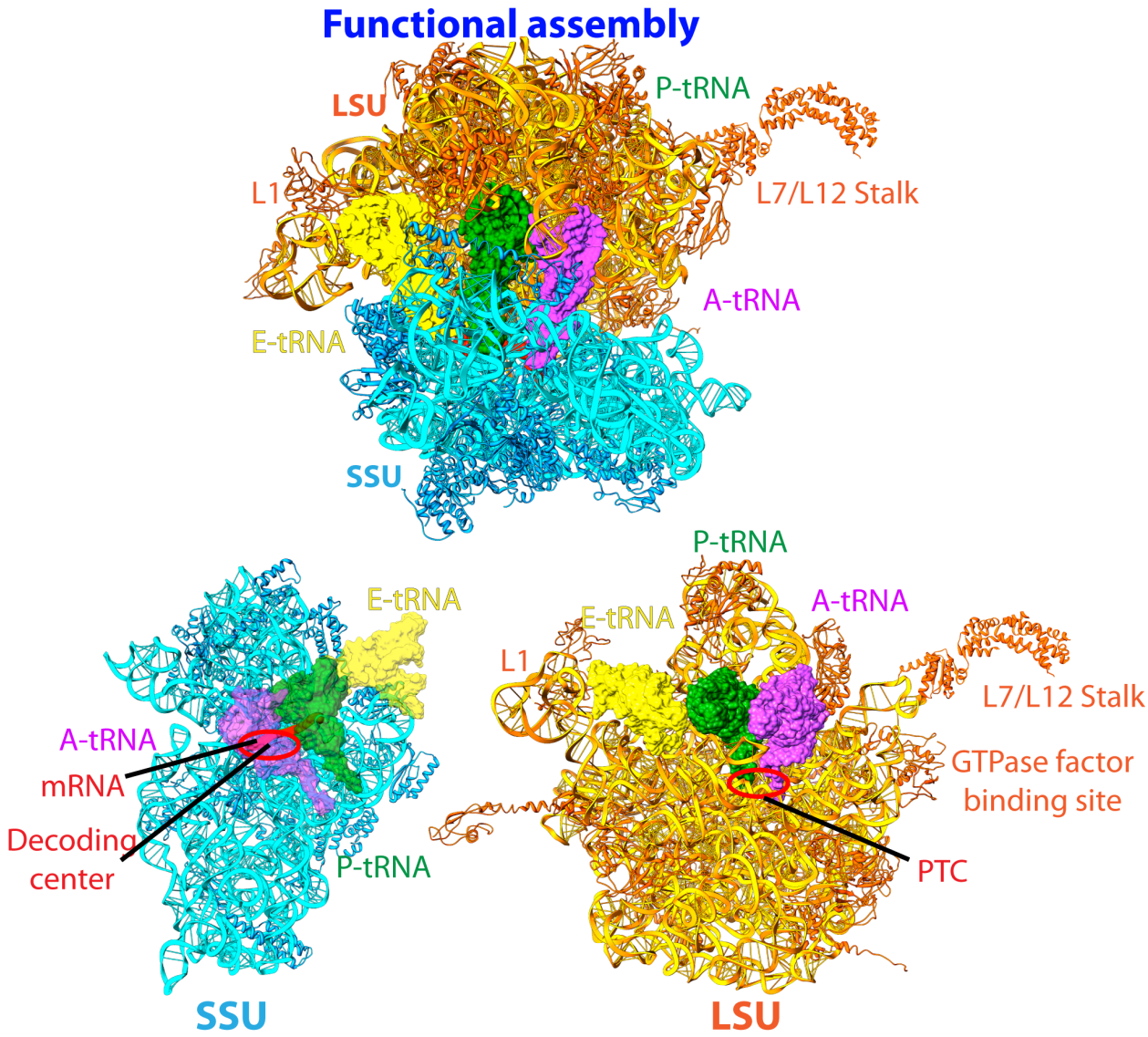


Figure 2.1: The structure of the 70S ribosome. Top panel is the 'Top' view of the 70S ribosome with mRNA (red) bound to an A-site (pink), P-site (green) and E-site (yellow) tRNA. Bottom panel is the 'interface view'. The decoding site in circled as is the PTC. rRNA is colored with lighter shade compared to the r-proteins. Figure rendered in UCSF Chimera with structures from rcsb.org, pdb id 2WDK and 2WDL. L7/L12 stalk was added from pdb id 1ZAW based on the model from Schmeing *et al* [52].

with one large ribosomal RNA (rRNA) and more than 70 ribosomal proteins depending on the species (their names suggest their centrifugal coefficient). In addition they contain 1-2 more small rRNAs (described below). All these molecules, aided by a host of other protein factors assemble into a complex (70S/80S) during translation [52]. Figure 2.1 is an overview of a bacterial ribosome with the different functional regions highlighted. The SSU mediates

interactions with mRNA and tRNA and sequentially decodes the genetic code in mRNA and the LSU catalyzes the peptide bond synthesis. Amino acids are brought into the reaction by amino acylated tRNAs (aa-tRNA), which in turn are delivered by protein factors. The ribosome has three binding sites for tRNAs: the A (aminoacyl) site for a new incoming aa-tRNA, the P (peptidyl) site that holds the growing peptide chain, and the E (exit) site through which the deacylated P-site tRNA exits after peptide bond synthesis (Figure 2.1 top). The SSU has the decoding site and also mRNA helicase activity (Figure 2.1 bottom). The LSU has the peptidyl transferase center (PTC) (Figure 2.1 bottom). In addition the LSU has a L7/L12 stalk, an extension of multimers of proteins L7/L12 that is required to bind GTPase factors in the GTPase center [53] and a L1 protuberance that controls tRNA exit from the E-site [54].

Translation in general proceeds in three phases, initiation, elongation and termination, although the specifics at each phase vary, particularly in the number of protein factors involved depending on the organismal superkingdom. Figure 2.2 provides an overview of the translation process in bacteria, which is better understood compared to other systems [52].

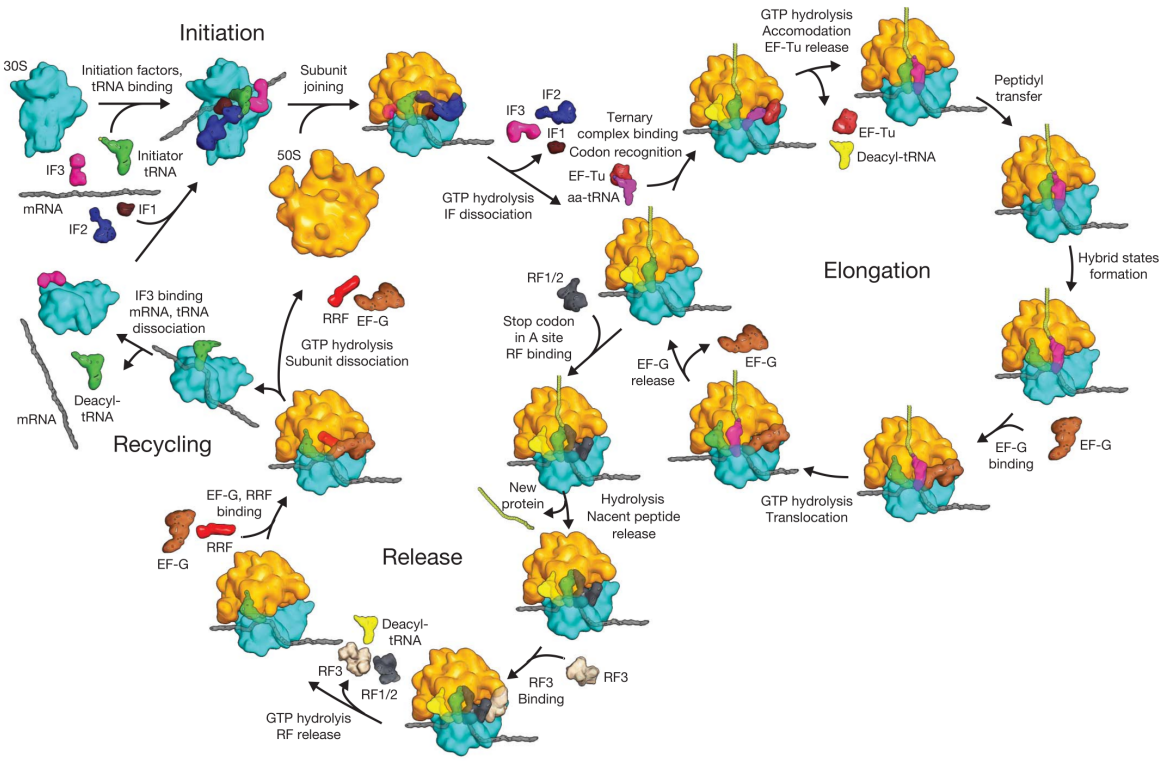


Figure 2.2: An overview of the bacterial translation process. Important intermediated stages are shown. LSU, SSU and tRNAs colored as in Figure 2.1. Figure from Schmeign *et al* [52] [©Nature Publishing Group]

Initiation: During initiation, an initiation complex is assembled when SSU binds mRNA at the start site of the coding sequence, precisely positioning the start codon in the P site, along with three initiation factors (IF1, IF2, IF3) and a special initiator tRNA (f-met tRNA) that binds to the P site.

Elongation cycle: The elongation cycle consists of three major processes; mRNA decoding, peptidyl transfer, and mRNA-tRNA translocation. (1) During decoding, the correct aa-tRNA is delivered to the A site by elongation factor Tu complexed with GTP (EF-Tu-aa-tRNA^{fMet}-GTP) a GTPase protein. Incoming tRNA is selected based on the codon-anticodon base pairing in the A site. Binding of the tRNA causes hydrolysis of GTP by EF-Tu and release of the factor from the ribosome through many conformation changes in their structures. These changes cause the aminoacyl end of the selected tRNA to propel into the PTC in the LSU where peptide bond formation occurs rapidly and spontaneously. (2) Following peptide bond formation, the A-site tRNA has a nascent peptide chain that is one residue longer and the P-site tRNA is deacylated. (3) During translocation, the SSU turns with respect to the LSU in a ratchet like motion when the P-site tRNA move to the E-site, and the A-site tRNA-peptide to the P-site. The hybrid state model, now proved experimentally shows that the 3′ ends of the A- and P-site tRNAs move first with respect to the LSU subunit to form a hybrid state of the ribosome, followed by movement of the mRNA and tRNAs with respect to the 30S subunit. This translocation requires the action of elongation factor G (EF-G), also a GTPase protein. This results in a free A-site with a new mRNA codon ready to select the next aa-tRNA.

Termination: When a stop codon reaches the A site, class I release factors (RF1 or RF2 in bacteria, eRF1 in eukaryotes) recognize the stop codon and catalyze the cleavage of the polypeptide chain from the P-site tRNA. Finally, a factor known as ribosome recycling factor (RRF), with the help of EF-G, disassembles the ribosome [52].

2.3 Secondary structure of rRNA

RNA structure in general is hierarchical [55]. The secondary structure is defined by the base-pairing interactions that act as a scaffold for higher order arrangements in three dimensions (3D). Secondary structures are composed of basic structural elements (substructures). The common

elements can easily be identified in the 5S rRNA molecule in Figure 2.3. Proceeding in the 5'to 3' direction of the sequence, there are 3 *helical stems* joined by a *multi-loop junction* at the center. Except for the helix that constitutes the ends of the RNA, all helices fold back onto themselves to form a *hairpin loop* of unpaired nucleotides. Unpaired segments that form *internal loops (bulges)* interrupt helical stems and are destabilizing in nature while the base pairing stabilizes the structure.

RNA constitutes the bulk of the ribosome [56]. Approximately two-thirds of cytoplasmic ribosomes are made up of RNA while proteins make up the remaining. In contrast, mitochondrial ribosomes are made up of more proteins than RNA, upto two-thirds. All SSUs have one large rRNA molecule termed 16S in Archaea and Bacteria (A/B), and 18S in Eukarya (E) [57]. All LSUs have a small 5S rRNA and a larger rRNA molecule termed 23S in Archaea and Bacteria, and 28S in Eukarya. Eukarya have an additional 5.8S rRNA that corresponds to the 5' end of 23S rRNA (Table 2.1)

Table 2.1: rRNA size and its proportion of the ribosome content in the three superkingdoms

Ribosome Source	Ribosomal RNAs (rRNA)
Bacterial, 70S, about 66% is RNA	SSU: 16S rRNA (~1500 nucleotides) and ~20 proteins LSU: 5S rRNA (~120 nucleotides), 23S rRNA (~2900 nucleotides) and ~30 proteins.
Archael, 70S, about 66% is RNA	SSU: 16S rRNA (~1500 nucleotides) and ~30 proteins LSU: 5S rRNA (~120 nucleotides), 23S rRNA (~2900 nucleotides) and ~40 proteins.
Eukaryotic 80S, about 60% is RNA	SSU: 18S rRNA (~1900 nucleotides) and ~30 proteins LSU: 5S rRNA (~120 nucleotides) 5.8S rRNA (~160 nucleotides), 28S rRNA (~4700 nucleotides) and ~45 proteins

The RNA and protein content of ribosomes from the tree superkingdoms of life are compared to describe the relative content of the respective macromolecule.

rRNAs are large, functional, non-protein-coding RNAs with a highly organized, hierarchical structure with distinct domains [58]. Since their functions are defined by their structures, rRNA structure is highly conserved. Although the primary sequence of rRNAs is generally conserved, the degree of conservation varies in different regions of individual rRNAs and has diverged significantly between Archaea, Bacteria and Eukarya. Despite significant

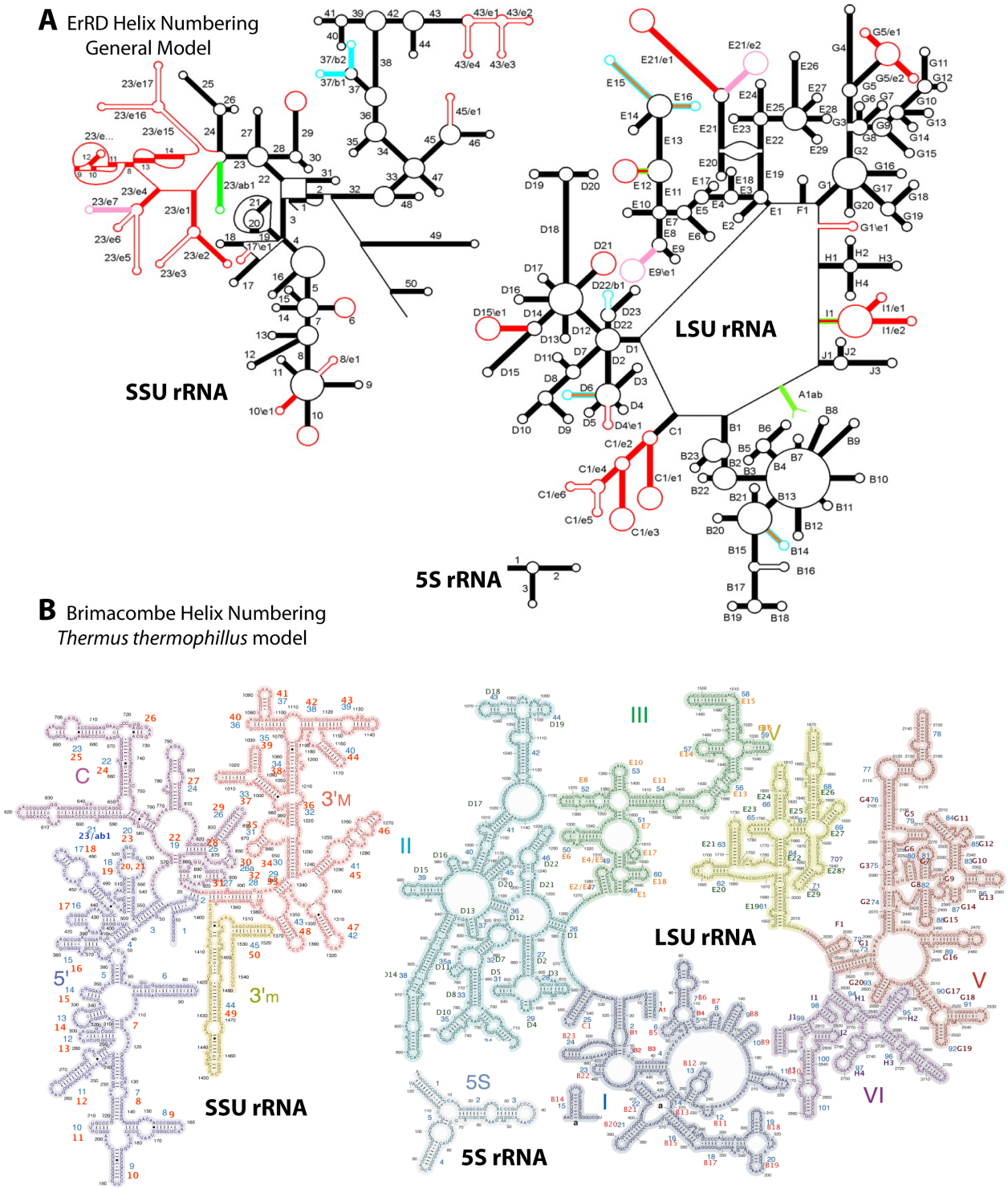


Figure 2.3: Secondary structure models of rRNA (A) General models of SSU and LSU rRNA secondary structure according to the ErRD. Helices are numbered starting from 5'-3' and number changes at every multi-loop junction. Color codes show in which species a helix is present; Black: All 3 superkingdoms; Red: Eukarya; Cyan: Bacteria; Green: Archaea; Orange: Archaea and Bacteria; Orange: Archaea and Eukarya; Solid lines: All species in a given group; Outlines: Subset of a group. (B) Secondary structures of *Thermus thermophillus* rRNA. Both ErRD and Brimacombe numbering are shown in different colors. RNA structure colors show the different domains.

differences in sequence both secondary and tertiary structures are remarkably conserved with interspersed distinctive features common to members within phyla of different superkingdoms [59]. Since secondary structure is determined by the base-pairing propensity of the primary sequence, sequence variation reflects secondary structure conservation by compensatory base changes in the corresponding paired regions forming double helical structures. This observation in the pre-crystallography era of the ribosome was in fact the basis for the predicted secondary structure model [60]. The first secondary structure model of the 16S rRNA, known as the covarion model was proposed in 1981 and proved to be accurate when the first high resolution crystal structure of the complete bacterial ribosome was solved in 2001 [61].

The secondary structure of SSU rRNA has ~50 helices and the LSU has about 100 helices that are conserved in all three superkingdoms [59]. In addition, there are many lineage specific insertions known as expansion segments. The structure is divided into domains based on the folding properties [56]. Figure 2.3 A shows the conserved and specific structural elements of rRNA secondary structure according to the European Ribosomal RNA Database (ErRD) models [62]. Universally conserved regions are in black. Figure 2.3 B illustrates the *Thermus thermophillus* model where the helices are defined by Brimacombe nomenclature [58, 63]. Both models follow similar comparative analysis and covariation models. However, the ErRD model is a generalized model to represent all possible sequences. Helices that are separated only by internal (or bulge) loops are considered as the same segment and those separated by multi-loop junctions are given different numbers [59]. The helix numbering also differs in terms domain of definitions but can easily be reconciled with the standard Brimacombe nomenclature as shown in Figure 2.3 B. The SSU is divided into 4 domains; the 5' Major, Central, 3' Major and 3' Minor domains and the LSU into 6 domains; domains I-VI and 5S rRNA is considered as domain VII.

2.4 Tertiary structure stabilizing interactions and motifs

RNA secondary structure specific topological constraints largely define global conformations and tertiary contacts act to stabilize specific conformations [64]. The folding of rRNA is further stabilized by r-proteins and divalent cations [65]. Many kinds of RNA-RNA and RNA-protein

tertiary interactions between the secondary structure motifs stabilize the complex 3D structure [66]. Most of these RNA tertiary structure motifs were initially identified in rRNAs and are well known. These include pseudoknots, tetraloops, and A-minor motifs. Psuedoknots are formed when the unpaired region in helix interacts with other loops/helices or forms a tertiary base pairing with another helix. A tetra loop is a specific type of a hairpin that limits conformation flexibility of the unpaired regions. Perhaps the most abundant of these motifs are A-minor interactions. A-minor interactions are extensive in rRNAs and are usually formed by highly conserved sets of nucleotides [66] where an unpaired stack of adenines lock into the minor groove of a helix and interact to pack the rRNA tightly. These motifs add to functionality and flexibility.

2.5 Approaches used to trace the evolution of rRNA

It is intuitive and probably obvious that such a large and complex RNA molecule could not have evolved *de novo* but gradually in many stages. Ever since the sequence and secondary structures of rRNAs were known, many models have been proposed to the possible ways rRNA genes and structures could have evolved [67-69]. Although some models are consistent with rRNA gene structure, genomic arrangement and expression or maturation pathways and are supported by comparative sequence analysis, most models are highly speculative. Initial sequence comparison studies showed that both SSU rRNAs and LSU rRNAs are conserved at the functional centers [70]. Based on the RNA world hypothesis it was proposed that the PTC evolved gradually from smaller RNAs [69, 71]. Since peptide synthesis is a spontaneous reaction when two amino acids are brought to close proximity and that simple RNA molecules could accomplish that by base-pairing rules, a peptide-synthesis-first origins for the ribosome in the LSU was proposed [71]. Due to the complexity of the structure it was also proposed that in the RNA world many smaller RNAs assembled together to synthesize peptide bonds. This proposal is supported by split rRNAs found in some unicellular eukaryotes and mitochondria. For example in *Euglena* the LSU is assembled from 14 different segments. However, it is hard to explain how, why and when a primitive PTC acquired the ability to decode genetic information [72, 73]

After crystal structures of ribosomes became available and a large body of biochemical experiments could be comprehensively integrated, and after specific roles of ribosomal components became clearer, models that use properties of the structural organization and tertiary interactions have been recently proposed [74-76]. These studies make inferences about the evolution of LSU rRNA and transitively about the ribosome itself, from tertiary structure [75] [76]. All these models explored only the LSU rRNA with a peptide-synthesis-first evolutionary rationale and hence proposed the origin of the rRNA in the PTC. Studies that analyze structure in a systematic way and within a comparative phylogenetic framework are few [41].

Phylogenetic methods have not only become central to reconstructing evolutionary history and testing hypotheses [77] but have also found practical applications in predicting and tracking epidemics [78]. Beyond biology, they have found applications in historical linguistics [24] and behavioral sciences [25]. Comparative morphology has long been the basis of phylogenetic inference, especially in paleontology [79] and is the only means to understand the evolutionary history of extinct species [80]. This approach was extended to molecular structures long ago and applied to organismal taxonomy [81] and more recently to describe the structural and evolutionary relationship of proteins [82]. More recently, evolutionary relationships were inferred on the basis of shared-derived rRNA structural features [41]. Although this study was a major advance both in terms of phylogenetic methods and opportunities to test hypotheses, it provided only a coarse grained tracing of the rRNA structure. A modification of this general methodological strategy has been used in this research to build on the initial findings, which has made it possible to decompose the rRNA secondary structure into its basic components and reconstruct a universal tree of structural components. The approach unifies phylogenetics and structural biology generating intrinsically rooted trees that ‘‘embed structure and function directly into phylogenetic analysis’’ [40]. The strategy is robust and has been employed in reconstructing deep rooted phylogenies of the living world [11, 12], uncovering reductive evolutionary tendencies in proteomes and a cellular origin for the tripartite world [83], tracing the origin and evolution of metabolic networks and proteins[84], exploring the origins of amino acid charging and the genetic code [85, 86], understanding the evolution of important functional RNAs including SINE RNAs [87], tRNA [88], 5S rRNA [89], and RNase P [89], and tracing

evolution of RNA structure in ribosomes [41]. This approach has made it possible for the first time, to provide phylogenetic support to theories of evolution of the translation apparatus.

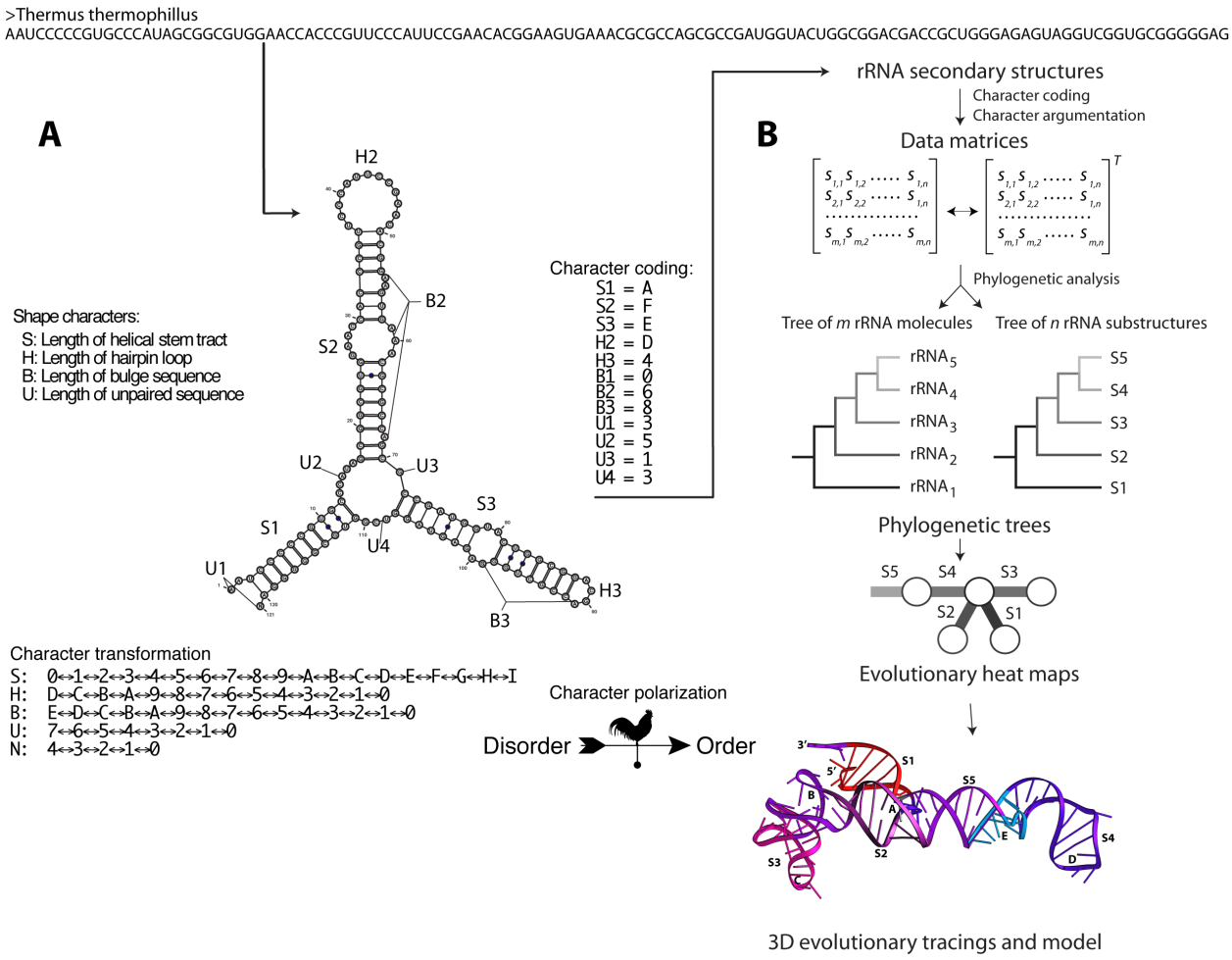


Figure 2.4: (A) Describes character coding and shows how 'shape' features describing RNA secondary structures can be used for tree reconstruction. Character states are treated as linearly ordered and polarized by invoking the thermodynamic propensity towards increased order to root the trees. (B) Once a matrix of character states is obtained it can used to reconstruct 2 kinds of trees. A tree of molecules; it represents a history of the molecules from different species over time. Another is the tree of the structural elements; it represents a history of how the structural elements evolved. After obtaining a tree the ancestry values are calculated and mapped on to the secondary structure and finally the 3D structure to produce an 'evolutionary heat map'.

The phylogenetic approach is illustrated here using the 5S rRNA structural elements that were described earlier. Figure 2.4 shows an overview of the character coding and tree reconstruction process. Since all cytoplasmic 5S rRNAs are ~ 120 bases in length and fold into a three-stem structure (Figure 2.4 A), structural elements such as stems (S), bulges (B), hairpin loops (H) and unpaired ends (U) are homologous. These decomposed homologous substructures

are equivalent to individual nucleotide or amino acid residues in a well-aligned set of homologous sequences. The characters defining the substructures can either be geometrical (shape) features or statistical (thermodynamic) features that describe a minimum free energy (mfe) structure. Only shape characters are considered in this research. Each character can be assigned a different quantitative character-state such as the number of paired nucleotides S. The variations in each state are pleisiomorphic (shared-derived) characters. Phylogenetic reconstruction yields a degree of relationship among different 5S rRNAs that are compared very much as in a sequence based tree. However, unlike sequence trees, the structural elements from different 5S rRNAs can also be compared to each other to reconstruct a tree of the structural elements. Remarkably, this tree provides a model of gradual accretion of structural elements and hence a model for the evolution of the structure itself. Since structure and function are related, the evolution of function related to the structures can also be deduced.

The method has been used to explore the origin and evolution of functional RNAs including SINE RNA, tRNA, 5S rRNA and RNase P. Analysis of tRNA structure showed the orgin of tRNA in the acceptor stem and supports the hypothesis that the top half domain composed of acceptor and pseudouridine (TΨC) arms is more ancient than the bottom half domain composed of dihydrouridine (DHU) and anticodon arms [90]. In addition, it corroborates the genomic tag hypothesis that postulates tRNAs were ancient telomeres in the hypothetical RNA world and their origins are in replication, not in translation [91]. Based on the trees a model for the evolution was proposed. According to the model, short hairpins homologous to the acceptor arm of extant tRNAs evolved with addition of helices homologous to TΨC and anticodon arms. The DHU arm was then added to form a proto-cloverleaf structure.

In a study of the evolution of the 5S rRNA it was found that 5S rRNA originated relatively fast but quite late during evolution at a time when a primordial translation apparatus and metabolic enzymes had already evolved [89]. A reconstruction of the tree of life resulted in a tripartite division rooted in Archaea. Together, the results led to the conclusion that 5S rRNA was incorporated late into the ribosomal ensemble that occurred prior to an (early) divergence of Archaea. This finding prompted us to exclude 5S rRNA in the current study and focus on the SSU and LSU rRNA.

RNase P, unlike tRNA or 5S rRNA, is a ribonucleoprotein complex (RNP) composed of the both protein and RNA. Combining results from the method explained above and a method that reconstructs evolution of protein structure from genomic census [92]. Results supported the early divergence of Archaea as inferred from 5S rRNA structure. In addition, importantly, it provided for a means to develop a model for the evolution of RNase P structure . This model shows that origins of RNase P are in structural elements that constitute the specificity domain that binds to its substrate tRNAs but not the catalytic domain. It also showed that RNA-protein interactions in catalytic RNP complexes are ancient. Comparison of the age of the RNase P proteins and other related proteins indicated that recruitment played a prominent role in the evolution of the catalytic domain of RNase P RNP complex. Enzyme recruitment in evolution of new function is a widely accepted hypothesis that proposes the promiscuous catalytic activities of enzymes provide a selective advantage to evolve new functions, which are recruited or co-opted into novel metabolic pathways [93].

2.6 Results and Discussion

The results explained below are from a combined analysis of both SSU and LSU rRNA structural elements in one tree. Unlike sequence methods, one of the advantages is that different parts of a multi-component complex can be analyzed simultaneously to deduce their interrelatedness. Phylogenies of structural elements were generated that are intrinsically rooted and provide a chronology of development of substructures (see Methods). Hence, the tree in itself becomes a model of structural evolution and can be used to deduce the relative age of different parts in complex molecular ensembles. As expected, functional centers were older in each subunit but there were also some unexpected results. Trees were also reconstructed separately for the SSU and LSU helices. The trees from individual subunits were congruent to the trees that were built from combined data. Results are confirmatory and are described separately in chapter 6: Appendix.

2.6.1 Evolution of the functional ribosomal core

Since ribosomal subunits can be considered a three dimensional (3D) arrangement of helices [94] and topological constraints of secondary structure greatly define global RNA structure [64], a universal tree of rRNA helices depicting the evolution of SSU and LSU rRNA secondary structure was reconstructed (Figure 2.5). A total of 93 sequences, 31 from each superkingdom was used to have a balanced sampling. To unfold data in trees, we calculated the relative age (ancestry) of each helix as a node distance (*nd*), the number of nodes from a hypothetical ancestor (root) in a relative 0-1 timescale (see Methods). These ages were traced in secondary and 3D structural representations of the molecules termed as "evolutionary heat maps" (Figure 2.5) and used to build timelines of development of components of the ribosome and their associated functions (Figure 2.6). The SSU rRNA is composed of 50 universal stem tracts (helices) arranged in four domains and the LSU rRNA has 100 universal helices arranged in six domains based on folding properties (Moore and Steitz, 2003). Only helices that are present in all three superkingdoms were included in the analysis.

Phylogenetic trees of combined SSU and LSU rRNA (Figure 2.5) show that SSU helix h44 is the oldest (*nd* = 0). This substructure, the penultimate helical stem in the SSU rRNA, is one of the most functionally important ribosomal substructures. It interacts with other SSU substructures responsible for mRNA decoding and with the LSU rRNA forming a functional relay (Cate et al., 1999). Most of the interactions of the mRNA and the tRNA are hence centered in this helix. This relay is proposed to link processes in the SSU decoding site with LSU based processes such as peptide bond formation and the release of elongation factors, thus modulating intersubunit interactions [95]. Helices h23, h24, h28, h30 and h34 are primordial (*nd* = 0.185-0.315); h23, h24, h28, h30 define the A, P and E sites of the SSU [95] and h34 is involved in tRNA translocation during the elongation cycle of translation [96]. However, some helices that are proximal to these ancient elements, such as helices h27, h29 and h31, are recent (*nd* = 0.444-0.722), suggesting they evolved after basic mechanisms were already established in the proto-ribosome, perhaps to refine established functions.

The LSU rRNA is twice the size of SSU rRNA and is divided into six domains (I – VI), 5S

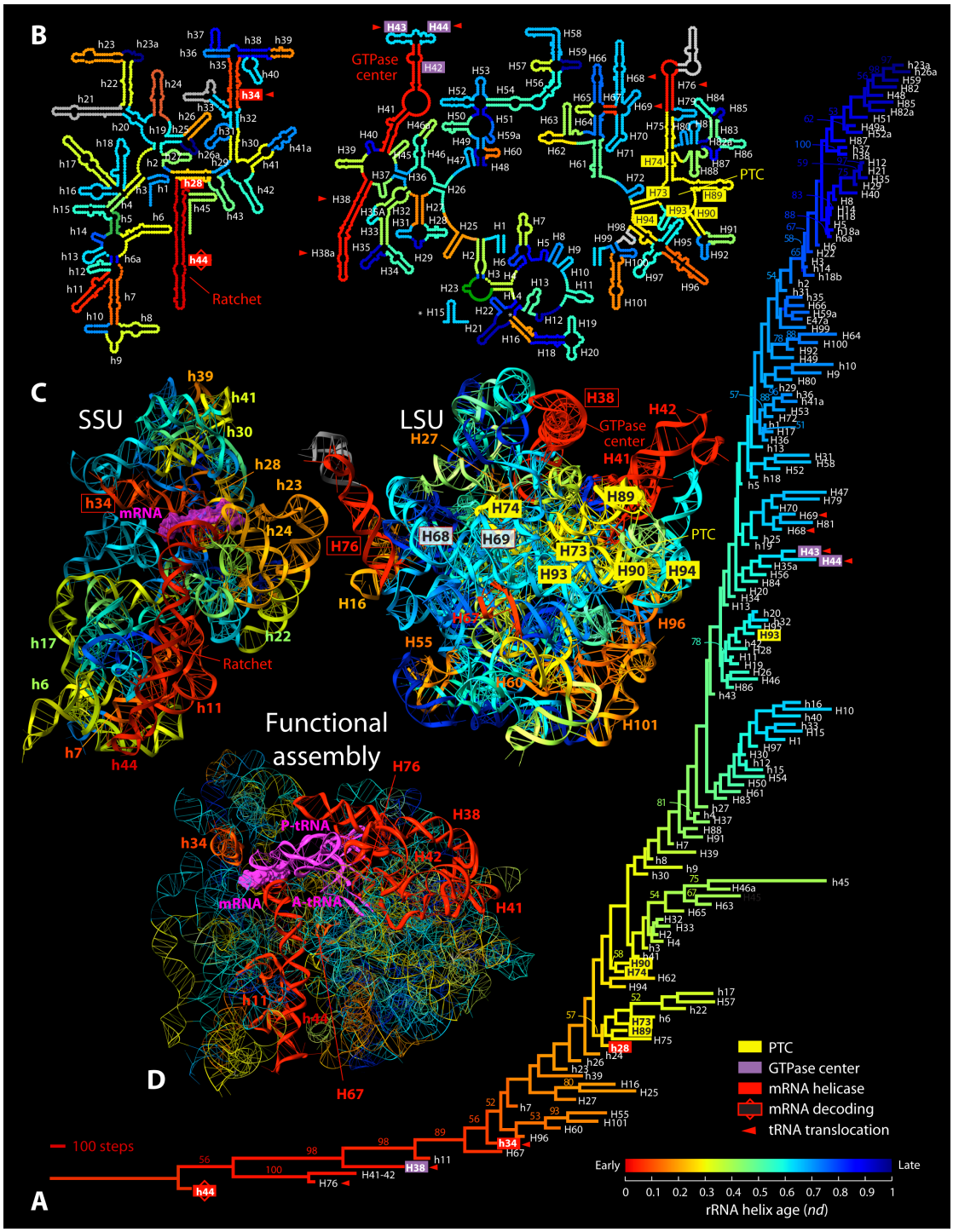


Figure 2.5: Relative age of rRNA structural elements. (A) Universal tree of SSU-LSU rRNA helical segments was reconstructed (33878 steps, CI = 0.168, RI = 0.710, HI = 0.831, g1 = -1.425) and relative age (*nd*) was determined. (B) SSU and LSU rRNA 2D 'evolutionary heat map' based on *nd* from (A). Functional centers are marked as highlighted the legend. SSU helix h44, which is the oldest component, is crucial functionally. LSU helices H38 and H41-42 which are required for the GTPase binding center and tRNA translocation are the next oldest, together they make the 'processivity center' (red). Interestingly the PTC developed later than the processivity region (yellow). (C) 3D 'evolutionary heat map'. (D) 70S ribosome with the oldest helices, namely those involved in processivity (red). The rapid appearance of the PTC is consistent with the ancient gene duplication theory proposed [97].

rRNA is considered the seventh domain [98]. The combined SSU and LSU rRNA (Figure 2.5) shows many functionally important regions are primordial. Helix H38 is one of the oldest substructures (*nd* = 0.037). It starts in the back of the particle, bends by about 90° and protrudes toward the SSU between domain V and 5S rRNA forming a crucial link between the two subunits [95]. Helices H73-H76, H89 and H90 that make up most of the catalytic core, the peptidyl transferase center (PTC) involved in peptide bond synthesis [99], are also ancient, (shaded yellow, *nd* = 0.296). The helical regions form the base of the polypeptide exit tunnel. In addition helices H2 and H7 of domain I (*nd* = 0.389), helices H26, H35, H35a, and H40 of domain II (*nd* = 0.6-0.9), helix H52 of domain III (*nd* = 0.648), and helices H61, H64, and H65 of domain IV (*nd* = 0.5-0.7) which are derived compared to helices of domain V, also comprise the peptide exit tunnel. Helices H32 and H69 that directly interact with the SSU are also derived (*nd* = 0.74). As with SSU, not all substructures that are proximal to the functional center are primordial or follow a serial chronology. Derived structural elements were therefore added to a basic functional proto-ribosomal unit later in evolution. This suggests the proto-ribosome was able to perform its function, perhaps less efficiently, with a simpler structure.

2.6.2 Early origins: A primitive processivity core precedes the PTC.

Figure 2.6 shows a timeline of accretion of the helical segments of the molecular ensemble and the emergence of functionally important regions for ribosomal processivity, namely the A-site, P-site and E-site, tRNA interactions, and intersubunit interactions. The timeline not only reveals concurrent structural diversification of the two subunits but it also uncovers the functional origins of the ribosome. SSU helices h44, h11, h34, and h7 (*nd* = 0.0-0.130), which are clustered around the most ancient h44 in 3D, are involved in mRNA and tRNA translocation. LSU helices H38, H41-42, H60, H67, and H96, which are less ancient (*nd* = 0.037-0.130), are all clustered around the PTC in 3D (Figure 2.5). Figure 2.6 F summarizes the start and end points of development of the core functions. It is clear that SSU helices harboring mRNA decoding, tRNA translocation and mRNA helicase activities precede the origin of LSU substructures that make up the PTC. In the core, h44, H38 and H67 together form more than half of inter-subunit bridge interactions. This is the processivity core of the ribosome. It performs the mechanically complex function of mRNA and tRNA binding and their translocation during the elongation

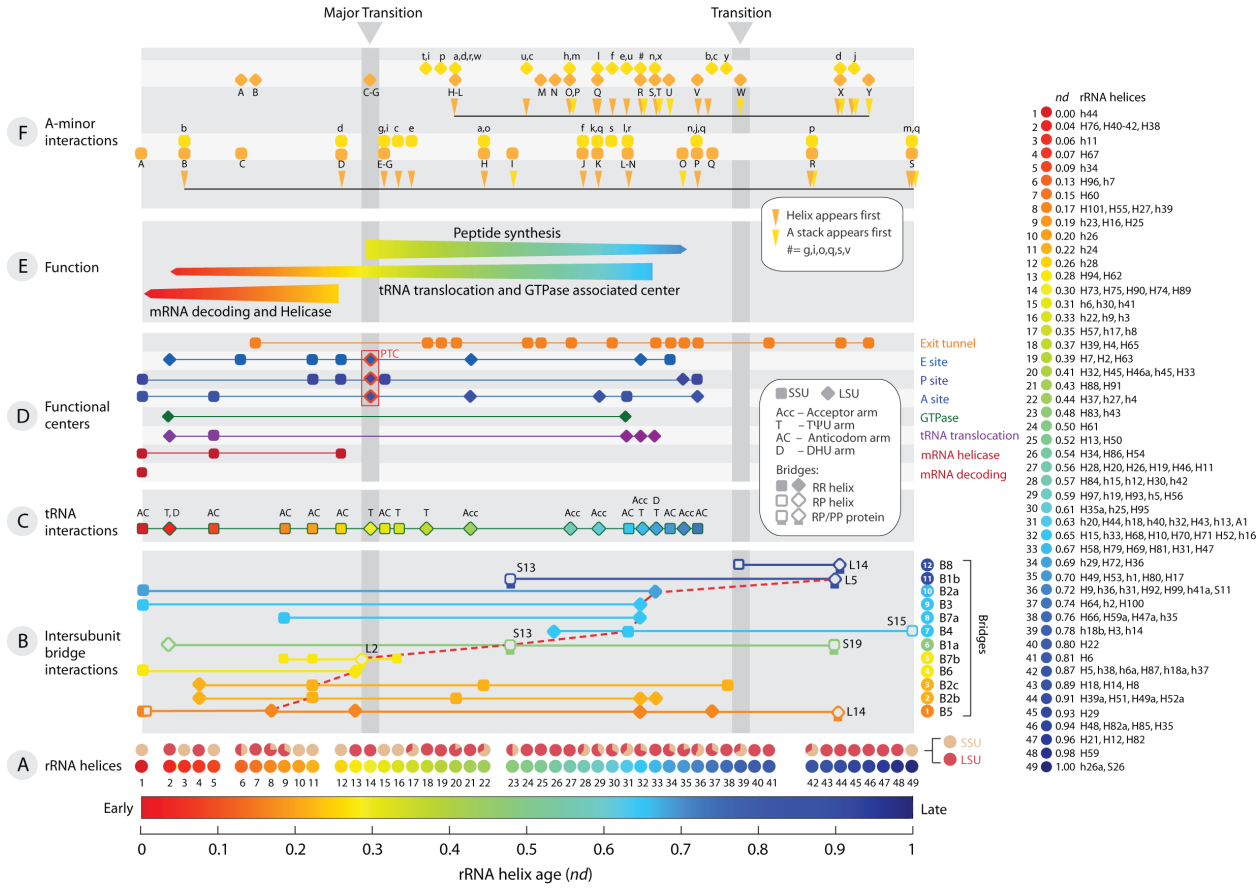


Figure 2.6.: Timeline of development of the different functional centers of the ribosome. The relative age of rRNA structural elements described by the tree in Figure 2.5-A. (A) The circles represent specific ages in the tree and are colored according to the ancestry values. Pie charts below each time point show the percentage of SSU and LSU substructures at each time point. The helices corresponding to each time point are listed on the right side. SSU helices have a prefix h and LSU helices H. The helices that form the different functional centers of the ribosome are indicated above each time point in different section (B)-(E). In these sections squares indicate SSU helices and rhomboids indicate LSU helices. (B) Intersubunit bridges, the age of a bridge is assigned as the age of first acceptor element of the donor-acceptor pair forming the bridge. (C) Helices that interact with tRNA; Acc-Acceptor stem, AC-Anti-Codon stem, T-TΨC stem, D-D stem. (D) Helices that form functional centers of the ribosome; purple is decoding center, green is the GTPase-associated center, blue is the A, P and E-site with the PTC highlighted in red box, orange is peptide exit tunnel. (E) Time points at which different functions started to develop. The width of the arrows shows the increase in number of helices forming the center and time taken for its development with the same colors as in A. (F) Shows the A-minor interactions; Names with the capital letters indicate the donor and small case indicates acceptor of the A-minor interaction.

cycle thus maintaining the reading frame and accuracy of translation. This is accomplished by a ratcheting action of the SSU relative to the LSU, and is driven by EF-G [100]. PTC functions are relatively simpler compared to ribosomal processivity.

Both proximity and orientation of tRNA substrates in the PTC are the sole driving force during peptide bond synthesis [101-104]. The PTC is accessible to the tRNA only after selection

by the decoding center mediated by elongation factor Tu (EF-Tu) [105]. Although the LSU can bind tRNAs by itself and synthesize peptide bonds with tRNA analogs the rates of peptide bond synthesis are orders of magnitude low [106]. *Peptide bond synthesis alone is not translation.* Full length tRNAs are required to achieve reaction rates equal to that of the LSU-SSU complex and to maintain the conformation of the PTC [107]. Correct selection of tRNA induces a signaling cascade affecting structural changes characteristic of allosteric mechanisms. Since SSU-facilitated selection of the correct aminoacyl-tRNA is the rate-limiting step for PTC activity we contend that the processivity center of the ribosome evolved before others to facilitate template directed polymerization.

Table 2.2: Showing the age of rRNA helicesage (nd) and the functions they are involved in.

Helix number		**Subunit**	**nd**	**Functional centers**				
Brimacombe	ErDB		(age)	mRNA decoding	mRNA helicase	tRNA translocation	GTPase center	PTC
h44	S49	SSU	0.000	+	+			
H38	D14	LSU	0.037			+		
H41	D17	LSU	0.037					
H42	D17	LSU	0.037				+	
H76	G4	LSU	0.037			+		
h11	S12	SSU	0.056					
H67	E25	LSU	0.074					
h34	S38	SSU	0.093		+	+		
h7	S8	SSU	0.130					
H96	H3	LSU	0.130					
H60	E18	LSU	0.148					
H101	J3	SSU	0.167					
H27	D2	LSU	0.167					
h39	S43	SSU	0.167					
H55	E12	LSU	0.167					
H16	B15	LSU	0.185					
h23	S25	SSU	0.185					
H25	C1	LSU	0.185					
h26	S29	SSU	0.204					
h24	S27	SSU	0.222					
h28	S32	SSU	0.259		+			
H62	E20	LSU	0.278					
H94	H1	LSU	0.278					+
H73	G1	LSU	0.296					+
H74	G2	LSU	0.296					+

Table 2.2 (cont.)

Helix number		Subunit	nd	Functional centers				
Brimacombe	ErDB		(age)	mRNA decoding	mRNA helicase	tRNA translocation	GTPase center	PTC
H75	G3	LSU	0.296					
H89	G16	LSU	0.296					+
H90	G17	LSU	0.296					+
h30	S34	SSU	0.315					
h41	S45	SSU	0.315					
h6	S6	SSU	0.315					
h22	S24	SSU	0.333					
h3	S3	SSU	0.333					
h9	S10	SSU	0.333					
h17	S18	SSU	0.352					
H57	E14	LSU	0.352					
h8	S9	SSU	0.352					
H39	D15	LSU	0.370					
H4	B3	LSU	0.370					
H65	E23	LSU	0.370					
H2	B1	LSU	0.389					
H7	B6	LSU	0.389					
H32	D7	LSU	0.407					
H33	D8	LSU	0.407					
H45	D20	LSU	0.407					
h45	S50	SSU	0.407					
H46a	D22	LSU	0.407					
H63	E21	LSU	0.407					
H88	G15	LSU	0.426					
H91	G18	LSU	0.426					
h27	S31	SSU	0.444					
H37	D13	LSU	0.444					
h4	S4	SSU	0.444					
h43	S48	SSU	0.481					
H83	G10	LSU	0.481					
H61	E19	LSU	0.500					
H13	B12	LSU	0.519					
H50	E6	LSU	0.519					
H34	D9	LSU	0.537					
H54	E11	LSU	0.537					
H86	G13	LSU	0.537					
H11	B10	LSU	0.556					
H19	B18	LSU	0.556					
H20	B19	LSU	0.556					
H26	D1	LSU	0.556					
H28	D3	LSU	0.556					

Table 2.2 (cont.)

Helix number		Subunit	nd	Functional centers				
Brimacombe	ErDB		(age)	mRNA decoding	mRNA helicase	tRNA translocation	GTPase center	PTC
H46	D21	LSU	0.556					
h12	S13	SSU	0.574					
h15	S16	SSU	0.574					
H31	D5	LSU	0.574					
h42	S47	SSU	0.574					
H84	G11	LSU	0.574					
h19	S22	SSU	0.593					
h5	S5	SSU	0.593					
H56	E13	LSU	0.593					
H93	G20	LSU	0.593					+
H97	H4	LSU	0.593					
h25	S28	SSU	0.611					
H35a	D11	LSU	0.611					
H95	H2	LSU	0.611					
H1	A1	LSU	0.630					
h13	S14	SSU	0.630					
h18	S19	SSU	0.630					
h20	S23	SSU	0.630					
h32	S36	SSU	0.630					
h40	S44	SSU	0.630					
H43	D18	LSU	0.630			+	+	
H44	D19	LSU	0.630			+	+	
H10	B9	LSU	0.648					
H15	B14	LSU	0.648					
h16	S17	SSU	0.648					
h33	S37	SSU	0.648					
H52	E8	LSU	0.648					
H68	E26	LSU	0.648			+		
H70	E28	LSU	0.648					
H71	E29	LSU	0.648					
H30	D6	LSU	0.667					
H47	E2	LSU	0.667					
H58	E15	LSU	0.667					
H69	E27	LSU	0.667			+		
H79	G5	LSU	0.667					
H81	G7	LSU	0.667					
h29	S33	SSU	0.685					
H36	D12	LSU	0.685					
H72	F1	LSU	0.685					
h1	S1	SSU	0.704					
H17	B16	LSU	0.704					

Table 2.2 (cont.)

Helix number		Subunit	nd	Functional centers				
Brimacombe	ErDB		(age)	mRNA decoding	mRNA helicase	tRNA translocation	GTPase center	PTC
H49	E4	LSU	0.704					
H53	E10	LSU	0.704					
H80	G6	LSU	0.704					
h10	S11	SSU	0.722					
h31	S35	SSU	0.722					
h36	S40	SSU	0.722					
h41a	S46	SSU	0.722					
H9	B8	LSU	0.722					
H92	G19	LSU	0.722					
H99	J1	LSU	0.722					
H100	J2	LSU	0.741					
h2	S2	SSU	0.741					
H64	E22	LSU	0.741					
h35	S39	SSU	0.759					
H47a	E3	LSU	0.759					
H59a	E17	LSU	0.759					
H66	E24	LSU	0.759					
h14	S15	SSU	0.778					
h18b	S21	SSU	0.778					
H3	B2	LSU	0.778					
H22	B21	LSU	0.796					
H6	B5	LSU	0.815					
h18a	S20	SSU	0.870					
h37	S41	SSU	0.870					
h38	S42	SSU	0.870					
H5	B4	LSU	0.870					
h6a	S7	SSU	0.870					
H87	G14	LSU	0.870					
H14	B13	LSU	0.889					
H18	B17	LSU	0.889					
H8	B7	LSU	0.889					
H40	D16	LSU	0.907					
H49a	E5	LSU	0.907					
H51	E7	LSU	0.907					
H52a	E9	LSU	0.907					
H29	D4	LSU	0.926					
H35	D10	LSU	0.944					
H48	E1	LSU	0.944					
H82a	G9	LSU	0.944					
H85	G12	LSU	0.944					
H12	B11	LSU	0.963					
H21	B20	LSU	0.963					

Table 2.2 (cont.)

Helix number		Subunit	nd	Functional centers				
Brimacombe	ErDB		(age)	mRNA decoding	mRNA helicase	tRNA translocation	GTPase center	PTC
H82	G8	LSU	0.963					
H59	E16	LSU	0.981					
h23a	S26	SSU	1.000					
h26a	S30	SSU	1.000					

Numbers in red are added to reconcile helix numbers not originally present

2.6.3 Intersubunit bridge history indicates early independent evolution of subunits.

The two subunits of the ribosome associate and communicate through intersubunit bridges and tRNAs. The subunit interface of SSU and LSU is almost devoid of proteins. The SSU subunit interface consists of three primordial rRNA helices, h44, h23 and h24, and one derived helix, h14. However, the LSU rRNA interface is made up of derived helices H68, H70, H71, H69 and H64, and two primordial helices H67 and H62. This is already strong indication that the two ribosomal subunits evolved independently before they interacted in modern translation. Figure 2.6 not only shows the timeline of development of functional centers but also roughly quantifies the number of substructures that make up the centers. The accretion of helices forming the processivity center occurs between *nd* = 0.0-0.3. In contrast, most helices forming the PTC appear together at *nd* ~ 0.3. The rapid and coordinated development of the PTC agrees with the proposal that it was formed by a duplication event [75, 97] and a self-folding ribosomal module [108]. Since the intersubunit bridge interactions hold the complex together we mapped these interactions to estimate when core ribosomal functions acted in concert.

Figure 2.6 shows the chronology of intersubunit bridge establishment. Bridge B5 is the oldest, first established between h44 and H27 (*nd* = 0.17) (Table 2.4). This initial bridge contact was immediately followed by the appearance of h24-mediated contacts in bridges B2b and B2c (*nd* = 0.22). These first three bridges involve the oldest SSU and LSU helices (helices h44, h24, H67 and H27; Table 2.4). Bridges B6 and B7b immediately follow, slightly preceding the establishment of the PTC (*nd* = 0.28-0.29). They also involve h44 and h24, but establish contacts with an ancient r-protein, L2. Bridge B1a was then established (*nd* = 0.48) and was

followed by the relatively quick appearance of bridges B4, B7a, B3, and B2a (*nd* = 0.63-0.67). Finally, B1b and B8 appear quite late in rRNA evolution (*nd* = 0.91). This progression of bridge interactions (red dotted line, Figure 2.6 B) corresponds to the gradual accretion of ribosomal substructures. Bridges B5, B2b, B7a, B3 and B2a form the functional core of intersubunit contacts. Mutations in any of these bridge contacts impair subunit association and translational fidelity [109]. Interestingly, one half of this functional core (B5, B2b) and roughly one half of all helices involved in bridge contacts (Table 2.3) originate concurrently with the processivity center

Table 2.3: The rRNA helices and r-proteins involved in the bridge interactions

			SSU		LSU	
Age	**Bridge**	**Type**	**rRNA**	**r-Protein**	**rRNA**	**r-Protein**
1	B5	R-R R-P	h44 h44		H27, H62, H64, H71	 L14
2	B2b	R-R	h24, h45		H67, H69, H71	
3	B2c	R-R	h24, h27		H66, H67	
4	B6	R-R R-P	h44 h44		H62	 L19
5	B7b	R-P	h22, h23, h24			L2
6	B1a	P-R		S13, S19	H38	
7	B4	R-R P-R	h20	 S15	H34 H34	
8	B7a	R-R	h23		H68	
9	B3	R-R	h44		H71	
10	B2a	R-R	h44		H69	
11	B1b	P-P		S13		L5
12	B8	R-P	h14			L14, L19*

R is RNA, P is protein, R-R is RNA-RNA bridge, R-P is RNA-Proteins interaction, h is SSU helix, H is LSU helix. L19 is not universal protein and not included in Figure 2.6 [110, 111].

of SSU and the other half of the functional core (B7a, B3, B2a) and all other bridges originate after the PTC.

The history of functions and interactions therefore suggests the two ribosomal subunits functioned at first independently and that a 'major transition' in evolution of translation at *nd* ~ 0.30 brought the two ribosomal subunits together into a protein biosynthetic ensemble.

2.6.4 Tertiary interactions increase after the first major transition.

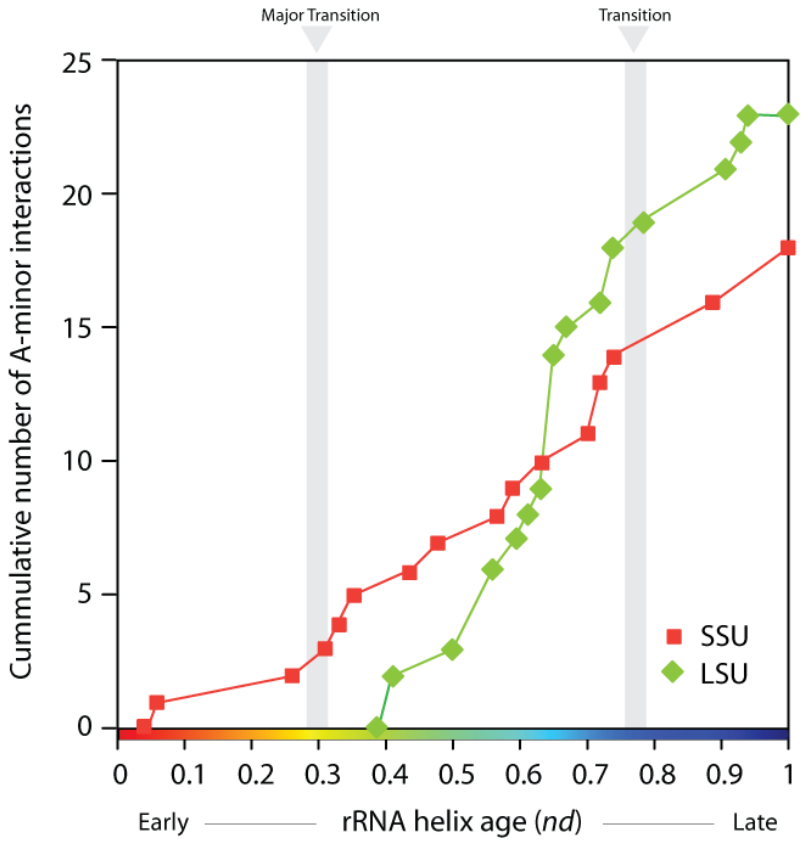


Figure 2.7: The increase in the number of A-minor interactions during the major transition. A sharp increase in the number of A-minor interactions allows for close packing of helices and greater stability during tRNA translocation.

The A-minor motif was first described after the crystal structure of the LSU rRNA was determined. A-minor interactions are extensive in rRNAs and usually formed by highly conserved sets of nucleotides [66]. In addition to stabilizing the rRNA structure, such interactions are also involved during decoding of mRNA [94]. The extent to which A-minor interactions are involved in ribosome function has prompted the study of their role in evolution

of the LSU rRNA on the assumption that the helix minor grooves into which adenosine stacks are inserted have evolved first [75]. We mapped all known A-minor interactions in both the SSU and LSU rRNA (Figure 2.6 F). The majority of the helices evolved before the corresponding adenosine stack. Interestingly > 90% of these interactions follow the first major transition, starting just after the development of the PTC and peaking immediately before the time of development of the GTPase associated center (Figure 2.6 F). During ratcheting motion associated with mRNA-tRNA translocation in the elongations cycle, very large conformational changes are required [112]. We propose that such interactions evolved to stabilize and maintain the ribosome structure during the elongation phase that led to high ribosomal processivity. Scarcity of such interactions before the major transition implies that the protoribsome structure was mostly stabilized by r-proteins. The increase in tight packing of helices and stability between the transitions is shown in (Figure. 2.7)

Table 2.4: The age of rRNA helices forming the A-minor interactions

LSU Helix		**nd**	**A-minor Stack**	**A-minor Helix**
ErDB	Brimacombe			
D14	H38	0.037		
D17	H41, H42	0.037	**u**	
G4	H76	0.037		
E25	H67	0.074		
H3	H96	0.13		**A**
E18	H60	0.148		**B**
D2	H27	0.167		
E12	H55	0.167		
J3	H101	0.167	**j**	
B15	H16	0.185		
C1	H25	0.185		
E20	H62	0.278		
H1	H94	0.278		
G1	H73	0.296		**C**
G2	H74	0.296		**F**
G3	H75	0.296		**G**
G16	H89	0.296		**D**

SSU Helix		**nd**	**A-minor Stack**	**A-minor Helix**
ErDB	Brimacombe			
S49	h44	0.000		**A**
S12	h11	0.056	**b**	**B**
S38	h34	0.093		
S8	h7	0.130		**C**
S43	h39	0.167		
S25	h23	0.185		
S29	h26	0.204		
S27	h24	0.222		
S32	h28	0.259	**d**	**D**
S6	h6	0.315		**E**
S34	h30	0.315	**g**	**F**
S45	h41	0.315	**i**	**G**
S3	h3	0.333		
S10	h9	0.333	**c**	
S24	h22	0.333		
S9	h8	0.352	**e**	
S18	h17	0.352		

Table 2.4 (cont.)

LSU Helix		nd	A-minor Stack	A-minor Helix
ErDB	Brimacombe			
G17	H90	0.296		**E**
E14	H57	0.352		
B3	H4	0.37		
D15	H39	0.37	**t**	
E23	H65	0.37	**i**	
B1	H2	0.389	**p**	
B6	H7	0.389		
D7	H32	0.407	**d**	**H**
D8	H33	0.407		**I**
D20	H45	0.407		
D22	H46a	0.407	**w**	
E21	H63	0.407	**a, r**	**J**
G15	H88	0.426		**K**
G18	H91	0.426		**L**
D13	H37	0.444		
G10	H83	0.481		
E19	H61	0.5	**u, c**	
B12	H13	0.519		**M**
E6	H50	0.519		
D9	H34	0.537		
E11	H54	0.537		
G13	H86	0.537		**N**
B10	H11	0.556	**h**	
B18	H19	0.556	**h**	**O**
B19	H20	0.556		
D1	H26	0.556		**P**
D3	H28	0.556	**m**	
D21	H46	0.556		
D6	H30	0.574		
G11	H84	0.574		
E13	H56	0.593		
G20	H93	0.593		**Q**
H4	H97	0.593	**l**	
D11	H35a	0.611	**f**	

SSU Helix		nd	A-minor Stack	A-minor Helix
ErDB	Brimacombe			
S50	h45	0.407		
S4	h4	0.444		**H**
S31	h27	0.444	**a, o**	
S48	h43	0.481		**I**
S13	h12	0.574		
S16	h15	0.574		
S47	h42	0.574	**f**	**J**
S5	h5	0.593		
S22	h19	0.593	**k, q**	**K**
S28	h25	0.611	**s**	
S14	h13	0.630		
S19	h18	0.630	**l**	**L**
S23	h20	0.630		**M**
S36	h32	0.630		**N**
S44	h40	0.630	**r**	
S17	h16	0.648		
S37	h33	0.648		
S33	h29	0.685		
S1	h1	0.704		**O**
S11	h10	0.722		
S35	h31	0.722	**n**	
S40	h36	0.722	**q**	**P**
S46	h41a	0.722	**j**	
S2	h2	0.741		**Q**
S39	h35	0.759		
S15	h14	0.778		
S21	h18b	0.778		
S7	h6a	0.870		
S20	h18a	0.870		
S41	h37	0.870	**p**	**R**
S42	h38	0.870		
S26	h23a	1.000	**m**	
S30	h26a	1.000	**m, q**	**S**
Sp Helix			**h**	

Table 2.4 (cont.)

LSU Helix		nd	A-minor Stack	A-minor Helix
ErDB	Brimacombe			
H2	H95	0.611		
A1	H1	0.63		
D18	H43	0.63	**e, u**	
D19	H44	0.63		
B9	H10	0.648	**s**	
B14	H15	0.648		
E8	H52	0.648	**o,i**	**R**
E26	H68	0.648	**g**	
E28	H70	0.648	**q, v**	
D5	H31	0.667	**n**	
E2	H47	0.667		
E15	H58	0.667		
E27	H69	0.667		
G5	H79	0.667		**S**
G7	H81	0.667	**x**	**T**
D12	H36	0.685		
F1	H72	0.685		**U**
B16	H17	0.704		
E4	H49	0.704		
E10	H53	0.704		
G6	H80	0.704		
B8	H9	0.722		
G19	H92	0.722		**V**
J1	H99	0.722		
E22	H64	0.741	**b**	
J2	H100	0.741	**c**	
E3	H47a	0.759		
E17	H59a	0.759	**y**	
E24	H66	0.759		
B2	H3	0.778		**W**
B21	H22	0.796		
B5	H6	0.815		
B4	H5	0.87		
G14	H87	0.87		

Table 2.4 (cont.)

LSU Helix		nd	A-minor Stack	A-minor Helix
ErDB	Brimacombe			
B7	H8	0.889		
B13	H14	0.889		
B17	H18	0.889		
D16	H40	0.907	d	X
E5	H49a	0.907		
E7	H51	0.907		
E9	H52a	0.907		
D4	H29	0.926	k	
D10	H35	0.944		Y
E1	H48	0.944		
G9	H82a	0.944		
G12	H85	0.944		
B11	H12	0.963		
B20	H21	0.963		
G8	H82	0.963		
E16	H59	0.981		

2.6.5 tRNA is at the center of ribosome evolution.

The proposed major transition corresponds not only to the rapid development of the PTC and bridges that link subunits but also to interactions with a full tRNA molecule in the A, P and E sites of the PTC (Figure 2.6 D). tRNAs have two structurally and functionally independent halves with independent evolutionary origins [88, 113]. The top half to which amino acid is charged contains the acceptor (Acc) arm and the TΨC arm (T), while the bottom half contains the anticodon (AC) and dihydrouridine (DHU) arm of the molecule. Each half of the tRNA interacts almost exclusively with one of the two ribosomal subunits [58], the top half with the LSU and the bottom half with the SSU. The development of the LSU rRNA helices of the PTC at *nd* = 0.30 added crucial contacts with the T arm of tRNA, and similarly several others appeared between *nd* = 0.30-0.37 (Figure 2.6 C). These events shifted the dynamics of contacts with tRNA. Before the major transition, most contacts involved ancient SSU helices and the

anticodon (AC) arm. After the transition, most contacts involved new LSU helices and the top half of the tRNA molecule. In fact, the first contact with the Acc arm occurred late, at *nd* = 0.37. This is remarkable. Most theories of ribosome evolution focus on peptidyl transfer and hence the PTC. However, the translation cycle consists of multiple processes that precede and follow the peptidyl transferase reaction in initiation, elongation and termination of protein synthesis [57]. The elongation cycle requires multiple, repetitive steps of tRNA selection and translocation involving conformational changes in the structure of LSU and SSU [114] and corresponding intersubunit movements with a ratchet like rotation of the SSU relative to the LSU [100]. The transferase reaction in the PTC is the simplest of all these process. The PTC simply positions the charged tRNAs in an optimal, proximal orientation with hardly any chemical contribution to enhance the rate of peptide bond synthesis [102, 103]. Instead, the 3' CCA end of Acc arm is critical for catalysis [115]. Thus tRNA is not a mere adapter as proposed by Francis Crick but an active (chemical) contributor to translation [116]. tRNA is central to and intimately involved in all steps of translation, including initiation, elongation and termination [117]. The 3' CCA end of the acceptor arm also contributes to fidelity at various steps of the elongation cycle [118] and is as critical to peptide release during termination as it is during peptide bond synthesis. Contacts with this Acc arm appear after the major transition, suggesting modern catalysis mediated by the 3' end of the Acc arm occurred well after the establishment of the PTC. That is, the PTC was not fully developed for accommodating specific tRNAs. The late appearance of H88 and H93 that are part of the PTC could explain this. Ribosomes and tRNAs possibly coevolved, transforming a simpler, error prone and sluggish primitive template directed polymerase process into a complex translation process that is accurate and faster. We propose the ribosomal complex was built around tRNA or tRNA-like structures (TLS) involved in replication that were later co-opted into a modern protein synthesis apparatus during a functional takeover.

The independent origins of the two functional halves of tRNA [88, 113], their almost exclusive interaction with only one rRNA subunit [58] and the independent history of these interactions (Figure 2.6 C) suggest rRNA subunits had originally different functions and were recruited for translation only after a modern cloverleaf tRNA-like molecule evolved. TLS involved in viral RNA replication [119], bacterial plasmid replication [120] and organellar DNA

replication [121] proposed to be relics of primitive tRNAs [91] support our hypothesis. Minihelices corresponding to the top-half (Acc arm) of tRNA can be substrates for ribosomal peptide synthesis [106] and EF-Tu [122]. Interestingly, the only tRNA arm that establishes interactions with the two subunits is the TΨC arm, and these interactions appeared at the time of the appearance of the PTC (Figure 2.6 C). This suggests the TΨC arm played a pivotal role during the major transition.

Interestingly the Acc arm is the most ancient part of the tRNA and the evolution of the tRNA structure suggests that the modern tRNA molecule had evolved much earlier than the modern ribosome [85, 86]. Although minihelices can mediate peptide bond formation the reaction rate is very slow [106], full sized tRNAs are required for reaction to proceed rapidly [107]. Thus we contend that modern translation evolved around tRNA by addition of new structural elements to a much simpler primitive ribosome. Mutational analyses of the newer rRNA and r-proteins structures do not abolish the functions but disrupt the efficiency of the translation process [123]. This corroborates the proposal that these structural components evolved to refine a less efficient system.

Table 2.5: Age of rRNA helices and their interactions with tRNA and other rRNA subunits

Helix number		Subunit	nd (age)	tRNA interactions				Bridge interactions
Brimacombe	ErDB			AC	T stem	D stem	Acc stem	
h44	S49	SSU	0.000	+				B5, B6, B3, B2a
H38	D14	LSU	0.037		+			
H41	D17	LSU	0.037					
H42	D17	LSU	0.037					
H76	G4	LSU	0.037					
h11	S12	SSU	0.056					
H67	E25	LSU	0.074					B2b, B2c
h34	S38	SSU	0.093	+				
h7	S8	SSU	0.130					
H96	H3	LSU	0.130					
H60	E18	LSU	0.148					
H101	J3	SSU	0.167					
H27	D2	LSU	0.167					B5
h39	S43	SSU	0.167					
H55	E12	LSU	0.167					
H16	B15	LSU	0.185					
h23	S25	SSU	0.185	+				B7b, B7a
H25	C1	LSU	0.185					
h26	S29	SSU	0.204					

Table 2.5 (cont.)

Helix number		Subunit	nd (age)	tRNA interactions				Bridge interactions
Brimacombe	ErDB			AC	T stem	D stem	Acc stem	
h24	S27	SSU	0.222	+	+	+		B7b
h28	S32	SSU	0.259	+				
H62	E20	LSU	0.278					B5, B6
H94	H1	LSU	0.278					
H73	G1	LSU	0.296					
H74	G2	LSU	0.296				+	
H75	G3	LSU	0.296				+	
H89	G16	LSU	0.296		+		+	
H90	G17	LSU	0.296					
h30	S34	SSU	0.315	+				
h41	S45	SSU	0.315					
h6	S6	SSU	0.315					
h22	S24	SSU	0.333					B7b
h3	S3	SSU	0.333					
h9	S10	SSU	0.333					
h17	S18	SSU	0.352					
H57	E14	LSU	0.352					
h8	S9	SSU	0.352					
H39	D15	LSU	0.370		+			
H4	B3	LSU	0.370					
H65	E23	LSU	0.370					
H2	B1	LSU	0.389					
H7	B6	LSU	0.389					
H32	D7	LSU	0.407					
H33	D8	LSU	0.407					
H45	D20	LSU	0.407					
h45	S50	SSU	0.407					B2b
H46a	D22	LSU	0.407					
H63	E21	LSU	0.407					
H88	G15	LSU	0.426					
H91	G18	LSU	0.426					
h27	S31	SSU	0.444					B2c
H37	D13	LSU	0.444					
h4	S4	SSU	0.444					
h43	S48	SSU	0.481					
H83	G10	LSU	0.481					
H61	E19	LSU	0.500					
H13	B12	LSU	0.519					
H50	E6	LSU	0.519					
H34	D9	LSU	0.537					B4
H54	E11	LSU	0.537					
H86	G13	LSU	0.537					

Table 2.5 (cont.)

Helix number		Subunit	nd (age)	tRNA interactions				Bridge interactions
Brimacombe	ErDB			AC	T stem	D stem	Acc stem	
H11	B10	LSU	0.556					
H19	B18	LSU	0.556					
H20	B19	LSU	0.556					
H26	D1	LSU	0.556					
H28	D3	LSU	0.556					
H46	D21	LSU	0.556					
h12	S13	SSU	0.574					
h15	S16	SSU	0.574					
H31	D5	LSU	0.574					
h42	S47	SSU	0.574					
H84	G11	LSU	0.574					
h19	S22	SSU	0.593					
h5	S5	SSU	0.593					
H56	E13	LSU	0.593					
H93	G20	LSU	0.593				+	
H97	H4	LSU	0.593					
h25	S28	SSU	0.611					
H35a	D11	LSU	0.611					
H95	H2	LSU	0.611					
H1	A1	LSU	0.630					
h13	S14	SSU	0.630					
h18	S19	SSU	0.630	+		+		
h20	S23	SSU	0.630					B4
h32	S36	SSU	0.630					
h40	S44	SSU	0.630					
H43	D18	LSU	0.630					
H44	D19	LSU	0.630					
H10	B9	LSU	0.648				+	
H15	B14	LSU	0.648					
h16	S17	SSU	0.648					
h33	S37	SSU	0.648					
H52	E8	LSU	0.648					
H68	E26	LSU	0.648				+	B7a
H70	E28	LSU	0.648					
H71	E29	LSU	0.648				+	B5, B2b, B3
H30	D6	LSU	0.667					
H47	E2	LSU	0.667					
H58	E15	LSU	0.667					
H69	E27	LSU	0.667			+		B2b, B2a
H79	G5	LSU	0.667		+	+		
H81	G7	LSU	0.667					
h29	S33	SSU	0.685	+	+			

Table 2.5 (cont.)

Helix number		Subunit	nd (age)	tRNA interactions				Bridge interactions
Brimacombe	ErDB			AC	T stem	D stem	Acc stem	
H36	D12	LSU	0.685					
H72	F1	LSU	0.685					
h1	S1	SSU	0.704					
H17	B16	LSU	0.704					
H49	E4	LSU	0.704					
H53	E10	LSU	0.704					
H80	G6	LSU	0.704					
h10	S11	SSU	0.722					
h31	S35	SSU	0.722	+				
h36	S40	SSU	0.722					
h41a	S46	SSU	0.722					
H9	B8	LSU	0.722					
H92	G19	LSU	0.722					
H99	J1	LSU	0.722					
H100	J2	LSU	0.741					
h2	S2	SSU	0.741					
H64	E22	LSU	0.741					B5
h35	S39	SSU	0.759					
H47a	E3	LSU	0.759					
H59a	E17	LSU	0.759					
H66	E24	LSU	0.759					B2c
h14	S15	SSU	0.778					B8
h18b	S21	SSU	0.778					
H3	B2	LSU	0.778					
H22	B21	LSU	0.796					
H6	B5	LSU	0.815					
h18a	S20	SSU	0.870					
h37	S41	SSU	0.870					
h38	S42	SSU	0.870					
H5	B4	LSU	0.870					
h6a	S7	SSU	0.870					
H87	G14	LSU	0.870					
H14	B13	LSU	0.889					
H18	B17	LSU	0.889					
H8	B7	LSU	0.889					
H40	D16	LSU	0.907					
H49a	E5	LSU	0.907					
H51	E7	LSU	0.907					
H52a	E9	LSU	0.907					
H29	D4	LSU	0.926					
H35	D10	LSU	0.944					
H48	E1	LSU	0.944					

Table 2.5 (cont.)

Helix number		Subunit	nd (age)	tRNA interactions				Bridge interactions
Brimacombe	ErDB			AC	T stem	D stem	Acc stem	
H82a	G9	LSU	0.944					
H85	G12	LSU	0.944					
H12	B11	LSU	0.963					
H21	B20	LSU	0.963					
H82	G8	LSU	0.963					
H59	E16	LSU	0.981					
h23a	S26	SSU	1.000					
h26a	S30	SSU	1.000					

2.7 Conclusions

Based on the results from this work and other previous works certain general conclusions could be made in addition to those that are concerned with the results presented in this chapter.

First, analyzing simple representations of molecular morphology within a phylogenetic framework, where a large body of theoretical and empirical evidence guides the assumptions, has proved to be a powerful approach to understand the evolution of biological functions [84, 85, 124, 125]. Structure based phylogenetic reconstruction methods have proven to be robust to analyze RNA secondary structures [90]. Results obtained here corroborate those findings. Furthermore, these results show that during evolution even if a function associated with the structure of a macromolecule is selected, the functional center is not always the origin of the macromolecule. Finally, structure-based phylogenetic reconstruction enables us to answer questions that sequence-based approaches fail to address due to the limitations with such methods which cannot be used to reconstruct deep-rooted trees.

The simultaneous analysis of both LSU and SSU rRNA structure now provides a means to determine which among them is relatively older. Although both the sequence of rRNA and proteins of the SSU are more conserved compared to the rRNA or proteins of the LSU, hypotheses about the origins of the ribosome have been usually centered on the PTC [74, 108]. Generally what is more conserved is considered more ancestral, but that reasoning is not favored when it comes to the ribosomal SSU. The argument against the SSU being older than LSU is

that the evolution of the genetic code would have not occurred if the protein synthesis mechanism had not evolved. This particular aspect will be addressed in a later chapter. However results presented show that the origins of protein synthesis are not in the PTC. Determining the relative age of different rRNA components show that a proto-ribosome was more likely to be a proto-SSU than a proto-LSU. Previous studies with the same methods have shown that in the case of RNase P, the origins of the molecule is not in its catalytic core [126] but in other helices that form the scaffold and stabilize the overall structure. The relatively late development of the PTC also shows a similar pattern of evolution in the ribosome although it is a much larger and more complex RNP. In addition, the evolution of tRNA has shown that the tRNA or tRNA like structures had evolved before the ribosomes and were involved in other processes including replication and probably metabolism. The results here show that both subunits evolved independently before tRNA mediated the subunit association.

In general results agree with previous findings that the functional core is relatively old. However, other studies are biased towards the LSU. Unlike other reports, the results described here are based on an objective criterion and provide unanticipated insights into the origins of ribosomal functions. The coalescence of many functions and patterns of evolution of the different structural elements agree with multiple lines of experimental and theoretical evidence. This cannot be coincidence. To our knowledge, this is the first time hypotheses of origins of the ribosome have been tested with standard phylogenetic methods.

Although protein synthesis is a tremendously complex and coordinated process, many consider the 'main' function of the ribosome is peptide bond synthesis [127]. Accuracy and processivity is the hallmark of processive enzymes and are at least as important as the catalytic activity if not more. The results from this research emphasize that all aspects that are important for biological function need to be considered to better understand its evolution. The evidence that the processivity center of the ribosome precedes the catalytic center supports the hypothesis that translation evolved as a result of a functional takeover of a related preexisting function, perhaps replication [50]. Since gene-replication and gene-translation are intricately related to the evolution of the gene [128], the results described here show that it is highly likely that the ribosome was recruited to translation from replication. The mechanistic similarities of the

ribosomal protein synthesis process to other processive enzymes like DNA and RNA polymerases [129] further corroborates our interpretations.

2.8 Materials and Methods

2.8.1 Data retrieval

The sequences of LSU and SSU rRNA were obtained from the European Ribosomal RNA Database (ErRD) at (http://bioinformatics.psb.ugent.be/webtools/rRNA/) [62] in DCSE format in which the secondary structure of the rRNA is encoded in helix numbering lines for sets of alignments specific for Archaea, Bacteria or Eukarya (Appendix, Figure 6.5). Helix numbering lines identify the corresponding paired regions of each helix in the rRNA secondary structure. A total of ~ 600 LSU and ~ 20,000 SSU sequences were obtained. Since the database is biased towards bacterial sequences, a balanced set of 93 sequences of both LSU and SSU from 31 species (limited due to unavailability of Archaeal sequences) each from Archaea, Bacteria and Eukarya were first used to reconstruct trees and later all usable sequences were included. More than 200 partial sequences were excluded. Data were analyzed with three different samplings. First, data from the original study with 35 sequences [130]. Second, data in this study with 93 sequences representing equal sampling from the three domains of life for which results are presented. Finally, complete set of sequences from the database (SSU: 19,184 and LSU: 593).

2.8.2 Determining relative age of rRNA structural elements

Since there are no explicit models for the evolution of RNA structure we limited our analysis to parsimony based methods implemented in PAUP* [131]. A novel phylogenetic method that reconstructs the history of molecular substructures of rRNA was developed earlier [11]. This method embeds structure directly into phylogenetic analysis [40]. Phylogenetic relationships were inferred on the basis of shared-derived characteristics in RNA structure using cladistic principles [130, 132]. Molecules were characterized by attributes that describe the topology of folded conformations. RNA secondary structures were first characterized using attributes that describe the overall ‘‘shape’’ (geometry) of molecules [133]. These attributes were then treated as linearly ordered multi-state characters that were polarized by fixing the direction of

evolutionary transformation toward molecular order. These trees describe a finite molecular system in which the 'leaves' represent the individual structural components of the molecule.

2.8.3 Character coding of RNA structure

RNA secondary structures are most suitable to study evolutionary relationships [134]. RNA secondary structures inferred from comparative sequence analysis were decomposed into structural elements (substructures) and their features (such as the length of stem tracts) were characterized using an alphanumerical format for cladistic analysis. Homologous components were treated as discrete entities and analyzed with maximum parsimony methods. Other alternatives are possible. In related studies, structural elements were characterized by their thermodynamic stability measured using their minimum Gibbs free energy increments [88]. These values were treated as discrete characters for maximum parsimony analysis. Coded characters were based on the length and number of double-helical stem tracts (S), hairpin loops (H), bulge and interior loops (B), and unpaired sequences (U).

In this study, topographic correspondence was the main criterion for determining character homology. It should be noted that unpaired nucleotides could form unusual base pairings or establish non-covalent interactions. These interactions are involved in high-order three-dimensional motifs like tetraloops, pseudoknots, A-minor motifs that stabilize RNA tertiary and quaternary structures are not considered in the structural models of this study. Several coding schemes are possible, however, character argumentation employed here is simplistic. That is, character coding disregards information and implications of higher order structure coarse-graining its three-dimensional complexities into a simple framework of non-interacting helical segments and thus have avoided any bias to a given substructure. Our assumptions are corroborated by rRNA crystal structures. Nearly all of rRNA is helical or approximately helical, and RNA structure can effectively be considered a three-dimensional arrangement of helical elements [94]. Character coding relies however on correct prediction of secondary structure. Covariation based comparative sequence analysis has been successful in predicting structures with high accuracy of up to 96% [61]. Structural inaccuracies were therefore assumed not to be severe and were tolerated as systematic error, provided structures result from a same comparative sequence study or are folded using the same algorithm.

The coding of rRNA was based on secondary structure models for the large and small subunits inferred from sequences deposited in the ErRD and defined by comparative sequence analysis (Wuyts et al., 2004). The SSU model contains 50 universal stem tracts (S) and several double-helical segments specific for Eukarya. The LSU model contains 100 universal stem tracts and several other stems specific to certain taxa. As described earlier the ribosome is essentially an arrangement of double helical stems. Thus helices (S) present in all three super kingdoms were used for the analysis. Note that universal stem tracts in these models are defined as those segments separated by multibranched or pseudoknot loops and are identified by numbers ordered in the 5'-to-3' direction. Character states were limited to 64, the maximum number accepted by PAUP* (http://paup.csit.fsu.edu/paupfaq/faq.html), and were represented by the numbers ***0–9***, case sensitive alphabets ***A-Z*** and ***a-z*** and special characters **@** and ***&.*** Structural features with longer than 64 nucleotide lengths were given the maximum state (&), and if missing, the minimum state (0). Structural alignments listed characters describing the structure in the 5'-to-3' direction as it is read in the sequence, and for each sequence segment, in the order S, B, H, and U. Stem tracts were defined by two complementary sequence segments and characters (named by a number and its prime) to account for the difference in nucleotide number between stem and unpaired segments. Helix numbering of the rRNA stems as in ErRD [59, 62] was used in the character coding and tree reconstruction exercises SSU helices are numbered S1-S50 and LSU helices are numbered A-I corresponding to the different domains. This was then reconciled with the standard Brimacombe numbering [63] used in the crystal structure of *Thermus thermophillus* ribosome [58].

The method was initially applied to 35 rRNA molecules sampled from all the three organismal superkingdoms of life and later extended to 93 molecules with equal representation from each superkingdom (A, B and E). Finally it was used to analyze all available sequences form the ErDB (> 20,000 sequences). An in-house software module, MARTEN [135], was used to code characters from DCSE alignments and to generate executable files for PAUP*.

2.8.4 Character argumentation and assumptions

Multistate characters were ordered and polarized. Hypothetical ancestral molecules were chosen as those having maximum base pairing, order, and thermodynamic stability. Character attributes represent transformation pathways and hypotheses of relationship that are falsifiable and link character states to each other using basic evolutionary assumptions or axioms [136]. Phylogenetic analysis of RNA structure rests on a very simple model and on the auxiliary assumption that there is an evolutionary tendency towards order (hypothesis of polarization). This tendency may represent an accurate depiction of generalized trends, but the model may fail to explain exceptions and departures to the trend. Character evolution was based a model of character state transformation in which states corresponding to increased order were defined as being ancestral (plesiomorphic). This hypothesis of polarization towards order is based on a general trend in RNA structure evolution where molecules evolve towards uniqueness, greater stability and modularity [47, 137]. Although this is a falsifiable hypothesis, there is sufficient theoretical and experimental evidence to support these polarization trends:

Thermodynamics. The thermodynamic theory of evolution [42, 44] develops general principles that are applicable to biological systems of all hierarchies, ranging from molecular ensembles to ecosystems [45]. According to this theory, biological systems are self-organizing and tend to increase the order and complexity of the system by dissipating the disorder to their surroundings [43, 138]. These thermodynamic principles generalized to account for non-equilibrium conditions have experimentally verified a molecular tendency towards order and stability that drives biological change [139]. A large body of theoretical evidence that maps the structural repertoire of evolving RNA sequences from energetic and kinetic perspectives [134], with some important predictions confirmed experimentally [140, 141].

Molecular mechanics. Studies of extant and randomized RNA sequences have shown that molecular evolution enhances conformational order and minimizes frustration. Randomization of mono- and dinucleotides in single-stranded nucleic acids have been used to assess the effects of composition and order of nucleotides in the stability of folded nucleic acid molecules and uncover evolutionary processes acting on folding of DNA and RNA [142]. In experiments, extant evolved RNA molecules encoding complex, functional structural folds were compared to oligonucleotides corresponding to randomized counterparts [141]. Unlike evolved molecules,

arbitrary sequences were prone to having multiple competing conformations. In contrast to arbitrary proteins, which rarely fold into well-ordered structures [143], these arbitrary RNA sequences were however quite soluble and compact and appeared delimited by physicochemical constraints such as nucleotide composition that were inferred in previous computational studies [47].

Phylogenetics. Finally, tendencies towards molecular structural order and the hypothesis for rooting of trees have been experimentally verified by phylogentic congruence between trees generated from RNA sequence and those generated from structure [130, 132, 144], in addition to congruence between phylogenies generated from geometric and statistical characters[85, 87, 145]. Polarizing characters in the opposite direction resulted in trees that were less parsimonious and had topologies incompatible with conventional taxonomy. Additional studies in our group with focus on structure such as hypotheses of organismal origin derived from global trees of tRNA structures and constraint analysis [86] and phylogenies of proteomes derived from an analysis of protein structures in entire genomic complements [92] proved to be congruent and provide further indirect support to our hypothesis of polarization.

2.8.5 Phylogenetic analysis

The relative ancestry of rRNA structural elements were reconstructed using maximum parsimony methods in PAUP* v. 4.0-b10 [131]. The ANCSTATES command was invoked to define ancestral character states and polarity of character transformation. Phylogenetic trees were derived from heuristic searches using tree-bisection-reconnection (TBR) branch swapping and simple addition sequence. Phylogenetic reliability was tested by the nonparametric bootstrap method implemented using 5000 pseudoreplicates.

2.8.6 Evolutionary Heat Maps

To better visualize the relative age of the different elements of the ribosomal ensemble and to understand how the functions associated with these structural elements, secondary structures of rRNA and the 3D structure of the ribosome were painted with colors corresponding to their respective *nd* values. Secondary structure diagrams of *Thermus thermophillus* rRNA

corresponding to the crystal structure of the 70S ribosome (PDB id 1GIX and GIY) were obtained from the Noller Lab website at (http://rna.ucsc.edu/rnacenter/ribosome_images.html). A RGB color scale corresponding to the *nd* values 0-1 with an interval of 0.01 was produced in matplotlib [146] using scripts available at http://matplotlib.sourceforge.net/gallery.html. The secondary structure models were modified and colored according to the *nd* values (see below). Helix numbering from the ErRD was reconciled with the Brimacombe numbering scheme. The crystal structures of *Thermus thermophillus* 70S ribosome (PDB id 2WDK and 2WDL) were also colored according to corresponding *nd* values of the rRNA helices.

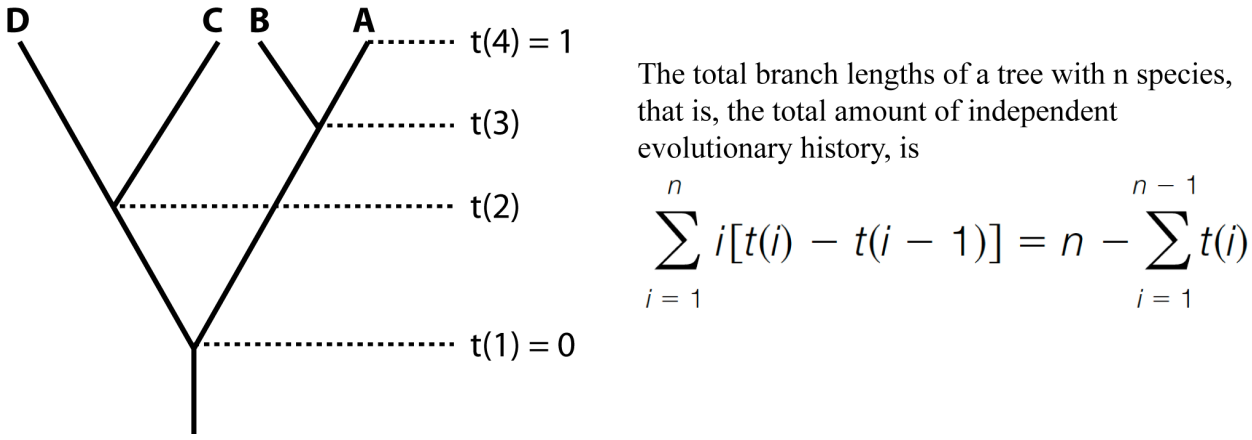


Figure 2.8: Node Distance (nd) is a measure of the relative age of the nodes on a tree. A small tree with cladogenic events ranging from a hypothetical ancestor at root to the present day is shown. The node distances are calculated as shown in the formula [147]

Calculating *nd* values: A *nd* value stands for '*node distance*'. Cladograms always bifurcate at a node (cladogenesis). A time scale is chosen such that the time from the first bifurcation to the last (present day) is 1, and *t(i)* denotes the time of the *i*th node [when the *i*+1 species appears; thus, *t*(*n*-1) is the time at which the nth species appeared, 1-t(*n*-1) time units ago]. For a tree with n species, t(n) is the present and t(n) ~1 regardless of the tree topology. In other words, it is the number of nodes (bifurcations) along the path to the taxon, divided by total number of nodes in the tree. A Perl script was used to calculate the *nd* values from the trees.

Chapter 3

Origin and evolution of ribosomal proteins

3.1 Introduction

Ribosomal proteins (r-proteins) play several important roles in ribosome function at every stage of translation [148]. Mutational analyses have shown that r-proteins contribute significantly to all the important functions of decoding, peptide synthesis and translocation [149]. In addition, r-proteins and multiple other protein factors are required during ribosome assembly for both directing and stabilizing ribosomal RNA (rRNA) folding [150]. The relative importance of the protein or RNA component of the ribosome in mediating the functions swayed one way or the other initially [149]. In the 1970s, it was agreed that r-proteins accounted for the functions of the ribosome and rRNA was simply a structural scaffold that held proteins in place for optimal positioning. Hence there was an elaborate search for a ‘ protein peptidyl transferase’ enzyme. After the search proved futile, proposals for a role of rRNAs in the function were made. By the 1980s the discovery of ribozymes (catalytic RNA) and biochemical data showing a role for RNA-RNA interaction to be crucial flipped the opinions to RNA centered functioning of the ribosome [151]. The high-resolution crystal structure of the LSU ribosome from an Archaea *Haloarcula marismortui* showed there were no proteins in the peptidyl transferase center (PTC) [99]. The crystal structure also showed that RNA forms the bulk of the structure and r-proteins generally occupy the peripheral regions of the ribosomal complex, away from the functional core at the intersubunit interface. Since then r-proteins have been attributed an auxiliary role in ribosome functions. In short, it was accepted that the ribosome is an RNA machine and the ‘ribosome is a ribozyme’. This was taken to be a major corroboration of the ‘RNA World’ hypothesis [152]. However, recent data have proved otherwise and attributed crucial roles for r-proteins (also). Biochemical experiments have shown that an r-protein was ‘at the heart’ of the

RNA machine and recent higher resolution structures have shown an important role of r-proteins in bacteria. It has also been pointed out that Archaea have a homolog of the PTC protein [153, 154]. In the LSU structure reported for *H. marismortui* it was disordered and hence unclear. This has raised the question if the ribosome is indeed a ribozyme [155].

3.2 Structure and function of r-proteins

The r-protein composition in ribosomes varies depending on the organismal superkingdom and also between species in each superkingdom [156]. Initial comparative sequence analysis of a few r-proteins available at the time showed that homologous r-proteins were highly conserved. Recently, analysis of ~70 fully sequenced genomes has provided a more comprehensive perspective [156]. Multiple conserved and superkingdom specific r-protein families have been identified with 68, 57 and 78 in Archaea, Bacteria and Eukarya respectively. Approximately 40% of these proteins are universally conserved. The nomenclature of r-proteins is according to the ribosomal subunit they are associated with, SSU r-protein families are named in a series starting with S1 and LSU r-proteins starting with L1. Figure 3.1 shows the distribution of r-proteins in the three superkingdoms. While ~40% proteins are conserved between the superkingdoms, there is very little sequence similarity between individual r-proteins.

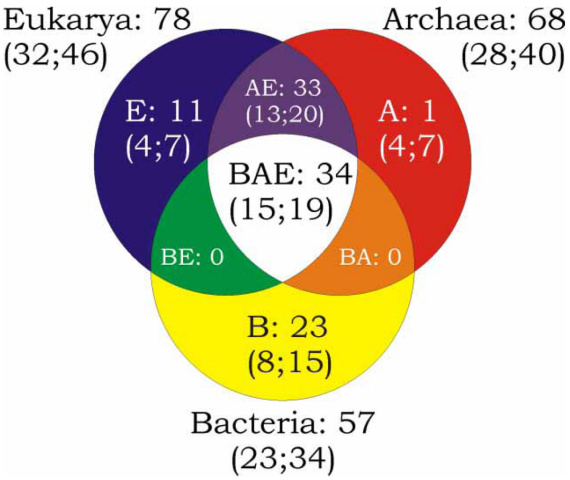


Figure 3.1: Distribution of r-proteins in the three superkingdoms. Numbers in parenthesis correspond to (SSU;LSU) respectively. Figure from [149].

The knowledge about the structure of r-proteins is mostly from bacterial and archaeal high-resolution crystal structures (~3.5 Å) . There is no crystal structure of a eukaryotic ribosome yet. However many protein structures have been deduced by homology models fit into cryo-electron microscopy maps (~8 Å). Figure 3.2 shows an overview of the r-protein structures and their position in the ribosome. Most r-proteins have unique folds; many unknown folds were described after crystal structures of r-proteins became available. Although at first glance it appears r-proteins occupy peripheral positions, many have long extensions that penetrate deep into the core of the functional centers. The extensions have a large compositional bias towards basic amino acid residues which neutralize the negatively charged RNA (phosphate) backbone at the core of the ribosome. The extensions are thus important for stabilizing the RNA and hence overall ribosome structure.

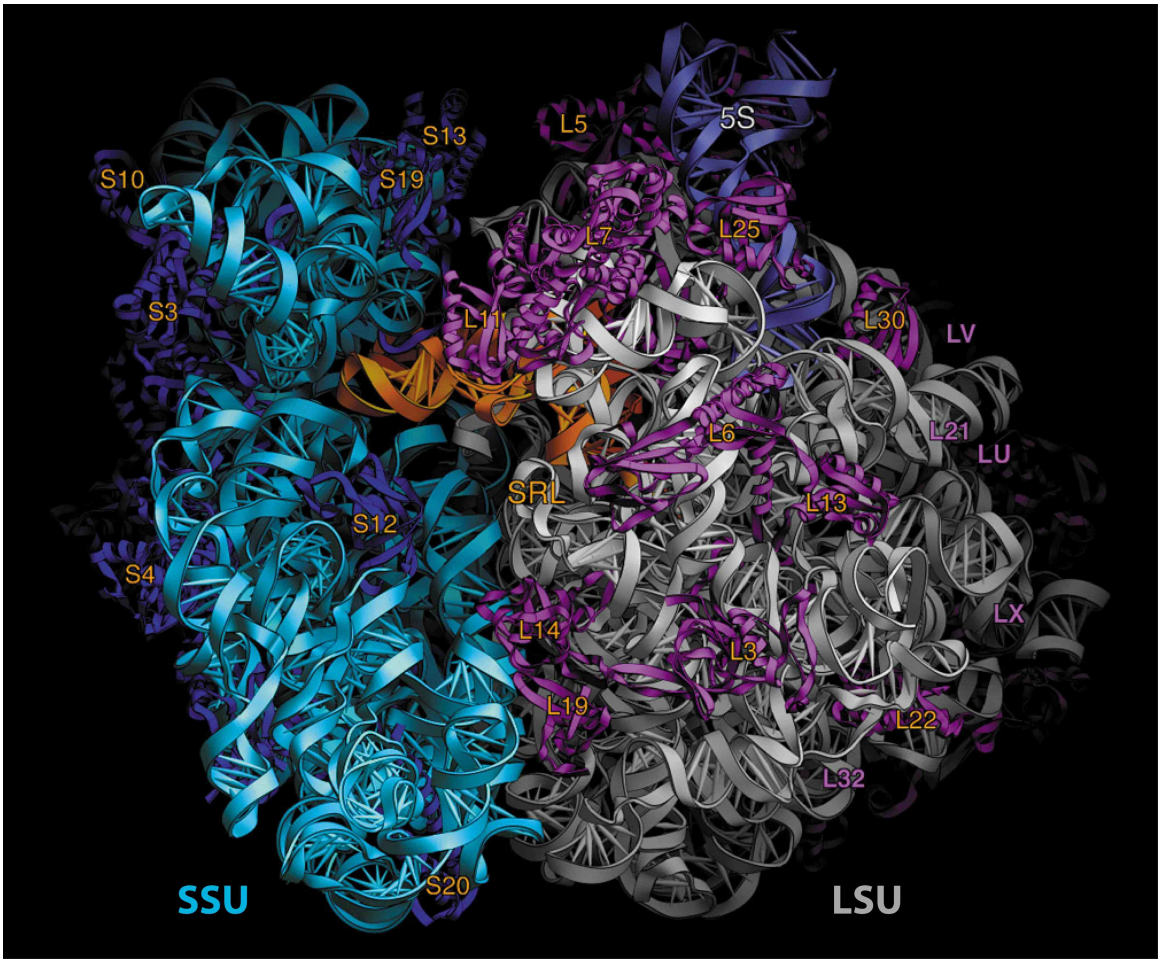


Figure 3.2: Structure of Ribosome showing the r-proteins. Grey; LSU rRNA, magenta; LSU r-proteins; dark blue; 5S rRNA; cyan; SSU rRNA; blue; SSU r-proteins. Orange; A-tRNA [157]

3.2.1 Ribosomal proteins and ribosome assembly

Ribosome assembly is a multi-step hierarchical process involving a host of extra ribosomal proteins in addition to r-proteins [150]. The assembly process is fairly conserved in all organisms but it is more complicated in the eukaryotes which posses a much larger rRNA and many more r-proteins compared to Archaea and Bacteria. Ribosome assembly starts with the transcription of rRNAs. They are transcribed as a single transcript with SSU, LSU and 5S rRNA, which is then subject to a series of posttranscriptional processes starting with excision into individual rRNAs by endonucleases. Following excision they are modified by methylation and psuedoridiylation. In addition many GTPases and ATPases aid the assembly process. Much of the knowledge about the assembly process is from in vitro reconstitution studies and predominantly of the bacterial ribosomal SSU. Preliminary results have suggested that the overall process in eukaryotes should be similar. However, eukaryotes have many more proteins, up to 200, with many found to be dispensable [158].

RNA and protein folding is hierarchical. Polymers first fold into their secondary structure motifs spontaneously before the tertiary interactions can be established. In RNA-protein (RNP) complexes, protein/RNA folding is highly cooperative. Generally, during RNA-protein binding, recognition almost always is by a 'induced fit' mechanism [159]. The free protein or RNA could either have disordered, flexible regions that require stabilization or attain new conformations upon binding that can alter functions or interaction with other molecules. During ribosome assembly both disorder minimization and stabilization occur, with r-proteins reducing the conformational entropy of the large rRNA secondary structures and globally stabilizing rRNA structures [160]. In addition, this initial folding is required for the secondary assembly proteins, which recognize these conformations for binding. The ribosomal assembly of the SSU has been better understood compared to the assembly of the LSU. In the SSU, the rRNA has a monolithic structure that largely determines the overall fold of the SSU. Although it can fold by itself into the known tertiary structure, especially in the central domain (Figure 2.3 B), it is not stable without r-proteins. The r-protein binding is not sequence-specific but is specific to the shape of the folded RNA. The induced-fit co-folding of the RNA and proteins is known to increase specificity of assembly and offset the energetic cost of folding regions that are disordered in the free protein or RNA. Thus, cooperative folding of rRNA and r-proteins

reduces conformational entropy and provides greater stability [160]. The three domains of SSU rRNA can be independently assembled as distinct RNP complexes. Furthermore, within the central domain a series of hierarchical conformational changes of RNA are followed by protein binding, with successive regions of RNA structure being stabilized by protein binding [150]. The hierarchical assembly map of the SSU is shown in (Figure 3.3).

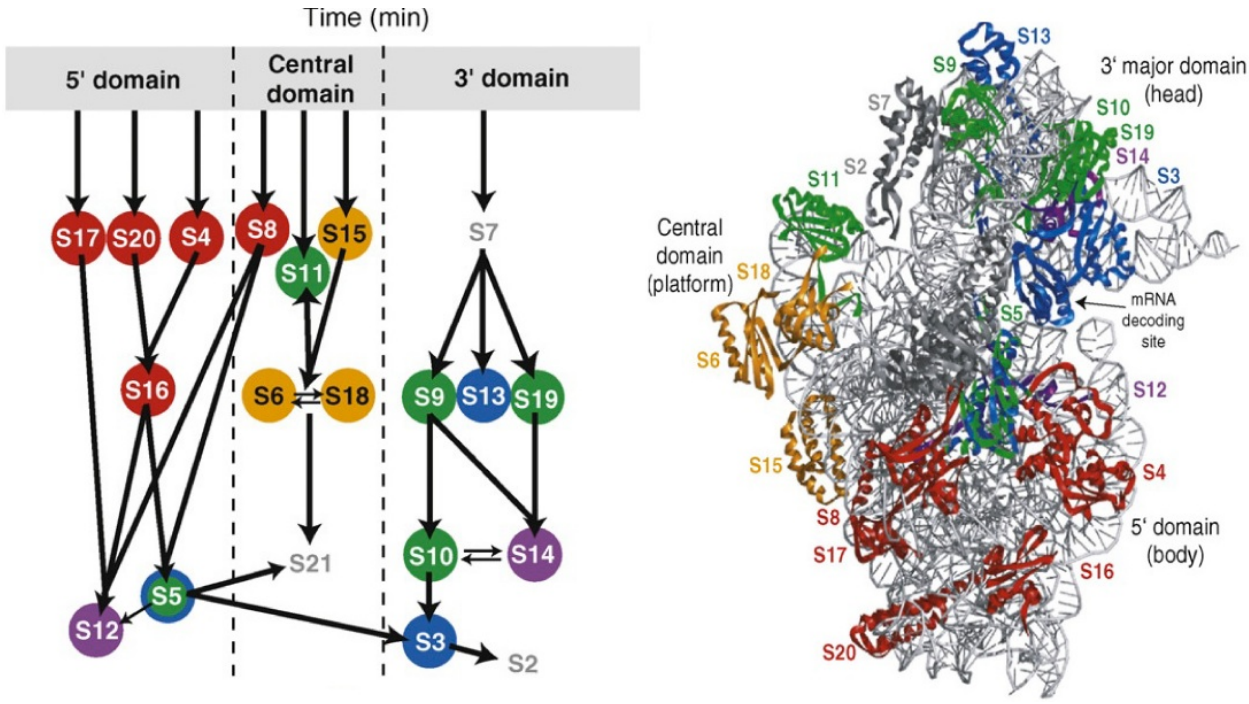


Figure 3.3: Kinetics of 30S assembly by pulse-chase mass spectrometry. Left, Nomura assembly map is colored by the protein binding rates at 37°C: red, >20 min^{-1}; orange, 8.1–15 min^{-1}; green, 1.2–2.2 min^{-1}; blue, 0.38–0.73 min^{-1}; purple, 0.18–0.26 min^{-1}. Arrows show hierarchical dependence. Right, 30S subunit from T. thermophilus, colored as in Left. Figure from [160]

The assembly of LSU is less understood compared to the assembly of SSU, but it is clear that it is more complex than that of the SSU as both the rRNA size and number of r-proteins is almost twice as that of the SSU [150, 158]. Unlike the SSU, the domains of LSU rRNA secondary structure do not correspond well to independent structural domains in 3D structure and this feature is thought to reflect a much higher degree of cooperativity (co-folding) between LSU rRNA and r-proteins.

In addition to the r-proteins many extra-ribosomal proteins are involved in the *in vivo* biogenesis of ribosomes. The number of proteins involved varies between the three

superkingdoms. In general these are found to be ATP and GTP-driven helicases and protein-chaperones. Helicases unwind folded RNAs and chaperones partially unfold proteins [158]. The role of these proteins implies an energy-dependent 'misfolding control' mechanism. Addition of some of these enzymes alleviates the required heating of the mixture during *in vitro* assembly of ribosomes.

The major difference between the archaeal/bacterial assembly and the eukaryotic assembly process is the large number of ATP and GTP-dependent enzymes. More than 200 are reported to be involved in eukarya [158]. This difference is attributed to intracellular translocation of ribosomes between compartments and a more complex spatial and temporal regulation of ribosome biogenesis. For example, the transport and binding of ribosomes to the endoplasmic reticulum requires many proteins that bind to r-proteins. When eukaryotic rRNA expansion segments and r-proteins are excluded from cryo-EM maps, the crystal structure of bacterial ribosomes can fit well into the map. This explains the additional complexity of eukaryotic ribosome structure and assembly.

In summary, the basic process of ribosome assembly is encoded in rRNA and r-protein sequences. A spontaneous, cooperative co-folding of the polymers initiates the assembly process, which is then driven and monitored by accessory factors and energy dependent enzymes.

3.2.2 Ribosomal proteins in ribosome function

The functional significance of r-proteins has only begun to be elucidated and is not exhaustive yet. As mentioned earlier, r-proteins contribute to every stage of protein synthesis. As expected, most of the proteins associated with functionally important sites are highly conserved. Due to the highly cooperative nature of ribosome function, neither can a specific protein be associated with a particular functional site nor can specific functions be assigned to individual r-proteins. The known and implicated functions of r-proteins are summarized in Table 3.1. Only a few important proteins associated in the decoding center, PTC and translocation activities are briefly described based on information from the detailed review in [149]

Decoding center: mRNA is threaded through the decoding center in the SSU through a compact tunnel composed of r-proteins S3, S4 and S5 and rRNA helices h28, h34 and h44. These proteins are found to be associated with mRNA helicase activity. Mutational analysis of the proteins either by changing amino acids or deleting 2-3 amino acids reduces the activity drastically [161]. S4, S5 and S12 are important for the accuracy of tRNA selection. Mutation or deletion of any one of these proteins increases the rate of mistranslation.

PTC: L16 and L27 are involved in the peptidyl transferase reaction by correctly orienting the A-tRNA and P-tRNA for catalysis to occur. L16 is required for peptidyl tranferase activity and for binding of the A-tRNA substrate. Deletion of L16 reduces the rate of peptidyl transfer. L16 has a homolog in Archaea, L10e, which also has a long extension and is predicted to have a similar role in the archaeal PTC [154]. L27 has a long N-terminal extension that penetrates into the PTC and interacts with the 3' ends of A-tRNA and P-tRNA and along with rRNA helices of the PTC, it orients the tRNA substrates correctly for catalysis. Deletion of 2-3 N-terminal amino acids reduces rate of peptidyl transfer. L27 is also involved in the late stages of ribosome assembly [162].

Translocation: The L1-protuberance and L7/L12-stalk are highly mobile elements of the LSU. L1 along with rRNA helix H77 is implicated in the release of E-tRNA during translocation. L7 is an N-terminal acetylated form of L12 and exists as a tetramer in *E. coli* and as hexamer in other species [$(L7/L12)^4$ or $(L7/L12)^6$]. The L7/L12 stalk consists of the multimer bound to a single copy of L10 and L11 and rRNA helices H42-H44. The L7/L12 stalk binds to EF-Tu and EF-G. Ribosomes depleted of L7/L12 complex can bind to EF-Tu or EF-G but cannot activate GTP hydrolysis by the factors. The L7/L12 complex is the only LSU r-protein that does not directly bind to rRNA.

3.2.3 Evolution of ribosomal proteins

Protein evolution is driven by many environmental and genetic factors. Large-scale genomic comparisons combined with ingenious experiments have shown that most of the amino acid replacements discovered so far significantly affect protein structural stability [163]. Unlike mutations that affect the functional centers, these mutations are spread all over the protein molecules reflecting the cooperative nature of protein folding. It also implies that structural

constraints dictating stability are among the strongest factors contributing to fitness. Random point mutations, gene duplication and divergence, and protein domain shuffling are among the predominant genetic mechanisms. Protein domains are fundamental structural, functional and evolutionary units defined by their folding properties [82].

Table 3.1: Functions associated with ribosomal proteins[a]

r-Protein	Functions
S1	Suggested to bring the mRNA into the proximity of the ribosome during initiation. Translational feedback regulation of S1 operon.
S3, S4 and S5	Form the mRNA entry pore and may have a helicase activity to unwind mRNA secondary structure encountered during translation.
S4	Mutations (ram) increase the error during the decoding process; role in rRNA transcription antitermination and translational feedback regulation of alpha operon.
S5	Probably facilitates changes of rRNA conformations that alters the selection mode of the ribosome from accurate to error prone and vice versa; mutations confer resistance against streptomycin and spectinomycin; ram mutations.
S12	Involved in decoding of the second and third codon positions at the A site. Mutations in S12 confer resistance against streptomycin, increase accuracy of the decoding process and, in most cases, concomitantly decrease the rate of translation. The lack of S12 in reconstituted particles also increases accuracy.
L1	Probably involved in the removal of deacylated tRNA from the E site. Translational feedback regulation of L11 operon.
L4	Mutations in L4 can confer resistance against macrolide antibiotics such as erythromycin by indirectly interfering with drug binding; role in rRNA transcription antitermination.
L7/L12	Involved in elongation-factor binding and GTPase activation. Together with L10, involved in translational feedback regulation of L10 operon.
L9	Mutations in L9 effect the efficiency of translational bypassing.
L11	Mutations in L11 or lack of the complete protein confer resistance against thiostrepton, an antibiotic that blocks the ribosomal transition from the pre-to post-translocational state and vice versa. During the stringent response this protein senses the presence of a deacylated tRNA in the A site; mutations or the absence of the protein can cause a relaxed phenotype (relC) resulting from loss of stringent control.
L16	May be involved in correct positioning of the acceptor stem of A-and P-site tRNAs as well as RRF on the ribosome. Mutations in L16 confer resistance to the orthomycins avilamycin and evernimicin. Homologue is L10e in archaea (and L10 in eukaryotes; Table 1).
L22	May interact with specific nascent chains to regulate translation. Furthermore, deletion of three amino acids in L22 confers erythromycin resistance without interfering with the binding of the drug.
L23	Present at the tunnel exit site and has been shown to be a component of the chaperone trigger factor binding site on the ribosome.
L27	Bacterial-specific protein implicated in the placement of the acceptor stem of P-site tRNA and binding of the ribosome recycling factor on the 50S subunit.
L29	Is located close to the tunnel exit site and may constitute part of the binding site for the signal recognition particle.

[a] Only ribosomal functions are listed. Some r-proteins are involved in extra-ribosomal functions not listed here

Despite immense sequence diversity proteins adopt relatively few folds, defined 3D arrangements of structure. The Structural Classification of Proteins (SCOP) database organizes proteins hierarchically into family, superfamily and fold based on sequence or structural evolutionary relationships [82]. A recent analysis explored the contributions of convergent and divergent evolution in the origin of protein folds. Remarkably it was discovered that not only did entire folds arise from an ancient prototype but some superfamilies from different folds from modular peptides of 20-40 residues. These results suggest that proteins may not have had many independent origins [164]. Most of the r-proteins are small single domain proteins and those that are larger are fusions of two domains. Hence, r-proteins are considered extremely ancient and their structures could provide a unique window into early protein evolution [165]. In fact it is proposed that many of the fold architectures of r-proteins could be the precursors of many folds. Very early genetic mechanisms like gene fusion, gene insertion and gene duplication is proposed to have pre-dated the development of the modern ribosome. There is no recognizable sequence homology between most of the r-proteins.

Due to their ancient history, protein structures are imprinted with deep evolutionary records [165]. A successful design becomes established by repeated usage. Hence folds that are abundant are likely to be older that others that are less frequently used. Hence the general phylogenetic approach described in the previous chapter can also be used to compare shared-derived features of fold abundance and fold usage to reconstruct a universal tree of protein architectures [166]. The approach described has been used to reconstruct trees both at the fold (F) and fold superfamily (FSF) levels. A description of how the method can be used to reconstruct trees of FSFs and determine the relative ages is found in [124]. Since the trees are intrinsically rooted, a timeline of appearance of different FSF can be determined. This approach is useful in determining the emergence and evolution of new functions associated with their corresponding FSFs. The approach has been used to determine global organismal phylogenies that reflect the conventional tripartite division of life at both fold and fold superfamily [12]. In addition, importantly the trees explained how the three superkingdoms emerged. Results from the study provided support for a communal ancestor that was complex as opposed to an ancestor with a minimal gene complement. Archaea were the first to diverge by losing a considerable repertoire of FSF architectures in response to adaptation to extreme environmental niches [124].

Similar to the approach used to reconstruct trees of rRNA structural elements, a modified matrix of the protein world can be used to reconstruct a tree of protein architectures. This tree provides for relative age of the protein Fs and FSFs. Not only will it provide for timing of structures but also a timing of evolution of different associated functions. This will test if rRNA or r-protein is more ancient. If not they could have coevolved, which is probably likely given the cooperativity of RNA and protein throughout the processes starting from the assembly.

3.3 Results and Discussion

3.3.1 Phylogenomics of protein structure reveals coevolution of r-proteins and rRNA.

r-proteins associate tightly with the ribosome, are extremely ancient, and their structures provide a unique window into early protein evolution [165]. The r-protein content varies in Archaea, Bacteria and Eukarya with 57, 68 and 78 r-proteins respectively; while 34 are universal, 11, 23 and 11 r-proteins are specific to the respective superkingdom [149]. To determine the relative evolutionary age of universal r-proteins we generated a universal phylogenomic tree that describes the evolution of protein domains at FSF level of structural complexity (Figure 3.5 A). The tree is rooted and was generated from a genomic census of FSF structures in 749 proteomes using established methodology (see Methods), and like rRNA substructure trees, it provides a timeline of appearance of proteins in the protein world [92, 167, 168]. The age of r-proteins (nd_P) (Table 3.2) was linked to the age of helices they contact (nd) to test the existence of coevolutionary patterns (Figure 3.4). Remarkably, the oldest r-proteins, S12 and S17 (nd_P = 0.018), interact with the oldest (h44) and second oldest (h11) SSU rRNA substructures, and remarkably, the linear correlation between the age of the most ancient rRNA contact (derived form the analysis of RNA structure) and the age of r-proteins (obtained from the census of domains in proteins) continues unabated until $nd \sim 0.35$ and $nd_P \sim 0.2$ (dashed lines, Figure 3.4 B). The correlation [$nd_P = -0.535\ nd + 0.009$; $R^2 = 0.961$; $F = 221.3$, $P < 0.0001$] was marked during early ribosomal history and strongly suggests both RNA and proteins co-evolve together as RNA-protein interactions establish with newly developed regions of the ribosome. The pattern of congruence also defines a general tendency that links protein and RNA timelines. We note that the early S12 and S17 proteins also interact with rRNA substructures h3, h4, h9 and

Table 3.2: r-protein nd and its corresponding SCOP superfamily.

SCOP ID	FSF	r-Protein	nd	nd norm	SCOP Superfamily
50249	b.40.4	**L2 -N**	0.018	0.00	Nucleic acid-binding proteins
50249	b.40.4	**S12**	0.018	0.00	Nucleic acid-binding proteins
50249	b.40.4	**S17**	0.018	0.00	Nucleic acid-binding proteins
54211	d.14.1	**S5 -C**	0.063	0.09	Ribosomal protein S5 domain 2-like
54211	d.14.1	**S9**	0.063	0.09	Ribosomal protein S5 domain 2-like
50447	b.43.3	**L3**	0.076	0.11	Translation proteins
50104	b.34.5	**L2 -C**	0.166	0.29	Translation proteins SH3-like domain
50104	b.34.5	**L19**	0.166	0.29	Translation proteins SH3-like domain
50104	b.34.5	**L24**	0.166	0.29	Translation proteins SH3-like domain
55174	d.66.1	**S4**	0.197	0.35	Alpha-L RNA-binding motif
54814	d.52.3	**S3 -N**	0.206	0.37	Prokaryotic type KH domain (KH-domain type II)
46946	a.156.1	**S13**	0.260	0.47	S13-like H2TH domain
56053	d.141.1	**L6**	0.269	0.49	Ribosomal protein L6
57716	g.39.1	**S14**	0.269	0.49	Glucocorticoid receptor-like (DNA-binding domain)
53137	c.55.4	**L18**	0.278	0.50	Translational machinery components
53137	c.55.4	**S11**	0.278	0.50	Translational machinery components
57829	g.41.8	**L32p**	0.283	0.51	Zn-binding ribosomal proteins
57829	g.41.8	**L33p**	0.283	0.51	Zn-binding ribosomal proteins
143800	d.325.1	**L28**	0.291	0.53	L28p-like
143800	d.325.1	**L31p**	0.291	0.53	L28p-like
54768	d.50.1	**S5 -N**	0.291	0.53	dsRNA-binding domain-like
55315	d.79.3	**L7ae**	0.296	0.54	L30e-like
55653	d.99.1	**L9 -C**	0.350	0.64	Ribosomal protein L9 C-domain
46992	a.7.6	**S20**	0.354	0.65	Ribosomal protein S20
74731	a.144.2	**L20**	0.381	0.70	Ribosomal protein L20
143034	d.301.1	**L35p**	0.395	0.73	L35p-like
48300	a.108.1	**L7**	0.417	0.77	Ribosomal protein L7/12, oligomerisation (N-terminal) domain
48300	a.108.1	**L12**	0.417	0.77	Ribosomal protein L7/12, oligomerisation (N-terminal) domain
57840	g.42.1	**L36**	0.417	0.77	Ribosomal protein L36
64263	d.188.1	**L17**	0.422	0.78	Prokaryotic ribosomal protein L17
54995	d.58.14	**S6**	0.422	0.78	Ribosomal protein S6
54565	d.27.1	**S16**	0.422	0.78	Ribosomal protein S16
55658	d.100.1	**L9 -N**	0.430	0.80	L9 N-domain-like
141091	b.155.1	**L21p**	0.430	0.80	L21p-like
46911	a.4.8	**S18**	0.435	0.81	Ribosomal protein S18
46561	a.2.2	**L29**	0.457	0.85	Ribosomal protein L29 (L29p)
54821	d.53.1	**S3 -C**	0.457	0.85	Ribosomal protein S3 C-terminal domain
47973	a.75.1	**S7**	0.475	0.89	Ribosomal protein S7
56047	d.140.1	**S8**	0.475	0.89	Ribosomal protein S8
55282	d.77.1	**L5**	0.480	0.90	RL5-like
52313	c.23.15	**S2**	0.480	0.90	Ribosomal protein S2
54570	d.28.1	**S19**	0.480	0.90	Ribosomal protein S19
50193	b.39.1	**L14**	0.489	0.91	Ribosomal protein L14
52161	c.21.1	**L13**	0.493	0.92	Ribosomal protein L13
54686	d.41.4	**L16p**	0.493	0.92	Ribosomal protein L16p/L10e
54999	d.58.15	**S10**	0.493	0.92	Ribosomal protein S10
46906	a.4.7	**L11 -C**	0.502	0.94	Ribosomal protein L11, C-terminal domain
54747	d.47.1	**L11 -N**	0.502	0.94	Ribosomal L11/L12e N-terminal domain
54843	d.55.1	**L22**	0.502	0.94	Ribosomal protein L22
52166	c.22.1	**L4**	0.507	0.95	Ribosomal protein L4
50715	b.53.1	**L25**	0.507	0.95	Ribosomal protein L25-like
56808	e.24.1	**L1**	0.516	0.97	Ribosomal protein L1
55129	d.59.1	**L30**	0.516	0.97	Ribosomal protein L30p/L7e
52080	c.12.1	**L15**	0.525	0.98	Ribosomal proteins L15p and L18e
54189	d.12.1	**L23**	0.529	0.99	Ribosomal proteins S24e, L23 and L15e
47060	a.16.1	**S15**	0.534	1.00	S15/NS1 RNA-binding domain
160369	d.58.62	**L10**	N/A		Ribosomal protein L10-like
64659	j.84.1	**L10**	N/A		Ribosomal protein L10
144321	j.118.1	**L34p**	N/A		Ribosomal protein L34p
58322	j.9.1	**S THX**	N/A		30S ribosomal protein THX

Universal proteins; yellow, LSU proteins blue, SSU proteins red, *nd*-norm is normalized to 0-1 scale for r-proteins (see 3.6.2)

h22 that are relatively derived (*nd* = 0.33-0.44). Similarly, many proteins start to interact with newer rRNA regions as they develop. Proteins appearing after the major transition also interact with older regions of rRNA, as new contacts involve also already established substructures. This indicates that r-protein precursors were interacting with the proto-ribosome very early and interactions continue to establish as the RNA structure evolve by accretion of new substructures (Figure. 3.4). Proteins also evolve interactions and possibly added new domains through indels. However, these very early peptide chains were synthesized by other means, perhaps through non-ribosomal peptide synthesis or as abiotic peptides [169] since modern ribosomal translation had not yet evolved. They likely helped stabilize ribosomal tertiary structure and caused structural rearrangements in RNA [170] enabling RNA structural conformations otherwise impossible by simple RNA-RNA interactions. These changes induced small improvements in translation speed and accuracy, providing strong selective advantages [127].

Although proteins constitute only a third of the mass of cytosolic ribosomes while RNA makes up the bulk [56], they contribute significantly to all stages of translation [148] and to rRNA assembly [160]. Biochemical studies of ribosomes depleted of several r-proteins and structural studies of the LSU that revealed absence of proteins in the PTC were used as evidence to suggest the ribosome is a ribozyme [99, 171]. Thus r-proteins were attributed only auxiliary roles in ribosome function. However, recent biochemical studies [153] and higher resolution structures of intact ribosomes with tRNA [154] have shown that r-protein L27 stabilizes P-site tRNA in the PTC and L16 facilitates aminoacyl-tRNA binding to the A site in bacteria. Mutations in these two proteins reduce the rate of peptidyl transfer. Although the r-proteins lie at the periphery of the ribosome and rRNA is mostly involved with main functions, many of them have extended tails that penetrate deep into the rRNA scaffold. These new revelations about r-proteins and catalytic mechanism of the ribosome with a predominant role of tRNA in substrate-assisted catalysis have raised doubts whether the ribosome is indeed a ribozyme [155]. The structural organization and stability of the PTC is most important for peptide bond synthesis. Ribosomal catalysis is thus a property of the integrated RNP complex rather than that of a confined section of RNA functional groups in the catalytic center. Both protein and RNA have

crucial roles that cannot be substituted with one another. We propose this complex functionality emerged from to the cooperative interaction of rRNA and r-proteins, which existed from the

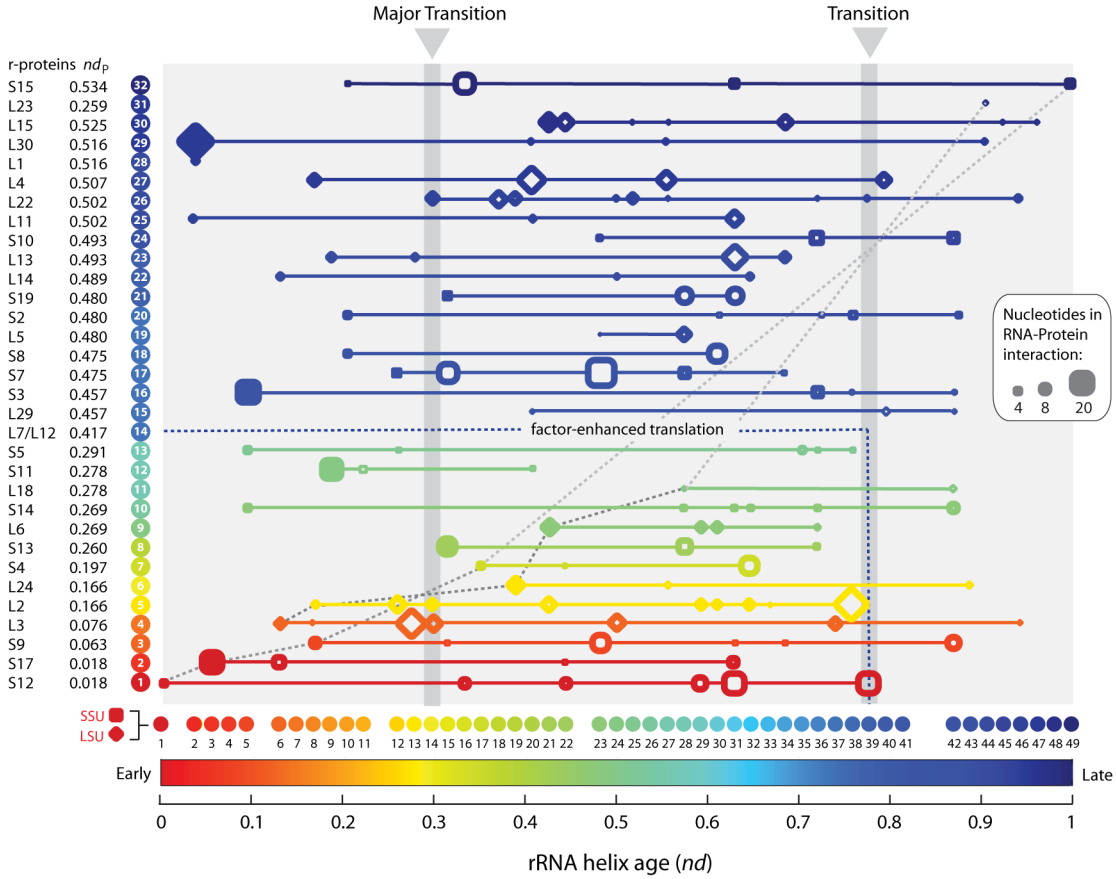


Figure 3.4: Tracing timline of rRNA helices establishing contacts with universal r-proteins. The relative age of the rRNA helices increases from left to right and r-proteins are ordered by age (from bottom to top) with corresponding nd_P value. The number of nucleotides at each time point involved in RNA-protein interactions is proportional to the squares (SSU) and rhomboids (LSU). r-proteins contacts are colored according to the age (nd) of the helix that makes the most ancient contact or is inferred from Figure**.3.5.**

earliest stages of ribosomal evolution and as rRNA coevolved with r-protein structure. Thus far *in vitro* peptidyl transferase activity catalyzed by protein-free rRNA derived from extant rRNA or ribozymes is not demonstrated [172]. Perhaps, the primordial cooperative property of the RNP complex explains why such attempts have failed.

Although rudimentary structures that could either catalyze peptide bond formation or interact with tRNA could have existed, unless the two act in concert it is not translation. So the existence of a peptide synthesizing apparatus need not be the sole precursor of a translation apparatus. For

example the precursors of present day non-ribosomal peptide synthetase enzymes [173] could have synthesized peptides.

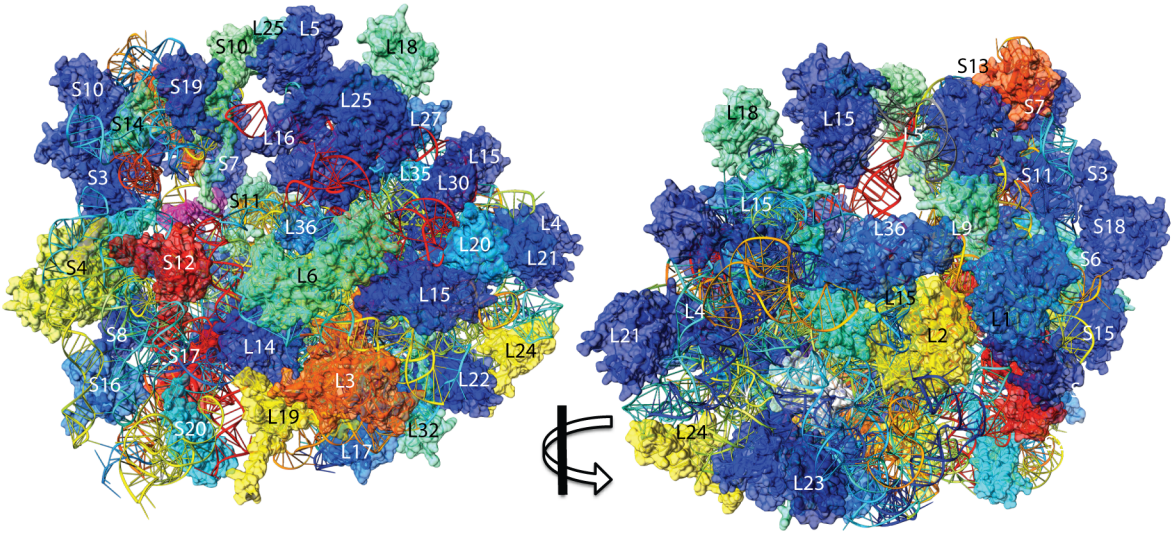


Figure 3.5: The age of r-proteins and their location in the functional assembly of the ribosome. Proteins are colored according to their age (nd). The right panel is 180-degree rotation of the left panel. Remarkably the relative age (nd) of the rRNA helices correspond to the relative age of their interacting r-proteins. The oldest proteins are associated at the intersubunit interface at the processivity center

Table 3.3: Age of rRNA helice interacting r-proteins. Interactions are quantified for comparison

FSF	r-Protein	nd r-protein	rRNA Contact		ndrRNA	Number bases interacting	
			Brimacombe	ErRD			
b.40.4	**S12**	0.018	h44	S49	0.000	2	47
			h3	S3	0.333	8	
			h27	S31	0.444	8	
			h19	S22	0.593	6	
			h5	S5	0.593	3	
			h18	S19	0.630	20	
b.40.5	**S17**	0.018	h11	S12	0.056	19	39
			h7	S8	0.130	10	
			h27	S31	0.444	2	
			h20	S23	0.593	7	
			h23/ab2	S23/ab2	N/A	1	
d.14.2	**S9**	0.063	h39	S43	0.167	8	54
			h30	S34	0.315	4	
			h41	S45	0.315	8	
			h43	S48	0.481	16	

Table 3.2 (cont)

FSF	r-Protein	nd r-protein	rRNA Contact		ndrRNA	Number bases interacting	
			Brimacombe	ErRD			
			h40	S44	0.630	3	
			h29	S33	0.685	4	
			h38	S42	0.870	11	
b.43.3	**L3**	0.076	H96	H3	0.130	12	92
			H101	J3	0.167	4	
			H94	H1	0.278	26	
			H73	G1	0.296	12	
			H90	G17	0.296	14	
			H61	E19	0.500	14	
			H100	J2	0.741	10	
b.34.4	**L2**	0.166	H74	G2	0.296	6	79
			H75	G3	0.296	8	
			H33	D8	0.407	11	
			H93	G20	0.593	8	
			H35a	D11	0.611	6	
			H67	E26	0.648	8	
			H79	G5	0.667	2	
			H66	E24	0.759	30	
b.34.5	**L24**	0.166	H7	B6	0.389	13	19
			H19	B19	0.556	6	
d.66.1	**S4**	0.197	h17	S18	0.352	9	50
			h5	S5	0.593	4	
			h18	S19	0.630	16	
			h16	S17	0.648	15	
			h23a	S26	1.000	3	
			h23/ab1	S23/ab1	N/A	3	
a.156.1	**S13**	0.260	h30	S34	0.315	15	32
			h41	S45	0.315	3	
			h42	S47	0.574	11	
			h31	S35	0.722	3	
d.141.1	**L6**	0.269	H91	G18	0.426	7	23
			H97	H4	0.593	8	
			H95	H2	0.611	8	
g.39.1	**S14**	0.269	h43	S48	0.481	4	29
			h42	S47	0.574	2	
			h32	S36	0.630	3	
			h31	S35	0.722	7	
			h36	S40	0.722	4	
			h38	S42	0.870	7	
			h37/b1	S37/b1		2	
c.55.4	**L18**	0.278	H38	D14	0.037	2	10

Table 3.2 (cont)

FSF	r-Protein	nd r-protein	rRNA Contact		ndrRNA	Number bases interacting	
			Brimacombe	ErRD			
			H85	G11	0.574	2	
			H87	G14	0.870	6	
c.55.4	S11	0.278	h23	S25	0.185	12	30
			h23	S25	0.185	9	
			h24	S27	0.222	6	
			h45	S50	0.407	3	
d.14.1	S5	0.291	h28	S32	0.259	4	21
			h1	S1	0.704	5	
			h36	S40	0.722	5	
			h35	S39	0.759	4	
			h35	S39	0.759	3	
c.21.1	L13	0.417	H25	C1	0.185	6	42
			H94	H1	0.278	6	
			H42	D18	0.630	22	
			H72	F1	0.685	8	
a.2.2	L29	0.457	H7	B6	0.389	13	13
d.53.1	S3	0.457	h34	S38	0.093	24	41
			h16	S17	0.648	1	
			h36	S40	0.722	7	
			h35	S39	0.759	4	
			h38	S42	0.870	3	
			h37	S41	0.870	2	
a.75.1	S7	0.475	h28	S32	0.259	8	62
			h30	S34	0.315	10	
			h41	S45	0.315	16	
			h43	S48	0.481	18	
			h42	S47	0.574	7	
			h29	S33	0.685	3	
d.140.1	S8	0.475	h26	S29	0.204	4	47
			h25	S28	0.611	15	
			h23/ab1	S23/ab1		28	
d.77.1	L5	0.480	H84	G11	0.574	16	16
c.23.15	S2	0.480	h34	S38	0.093	1	19
			h26	S29	0.204	2	
			h25	S28	0.611	6	
			h36	S40	0.722	3	
			h35	S39	0.759	2	
			h37	S41	0.870	5	
d.28.1	S19	0.480	h30	S34	0.315	10	41
			h42	S47	0.574	8	
			h32	S36	0.630	4	

Table 3.2 (cont)

FSF	r-Protein	nd r-protein	rRNA Contact		ndrRNA	Number bases interacting	
			Brimacombe	ErRD			
			h31	S35	0.722	6	
			h37/b1	S37/b1		13	
b.39.1	**L14**	0.489	H96	H3	0.130	4	24
			H71	E29	0.648	4	
			H92	G19	0.722	4	
			H61	E19	0.500	4	
			H95	H2	0.611	8	
d.41.4	**L16**	0.493	H89	G16	0.296	8	45
			H39	D15	0.370	14	
			H42	D18	0.630	8	
			H81	G7	0.667	4	
			H80	G6	0.704	5	
			H40	D16	0.907	6	
d.58.15	**S10**	0.493	h41a	S46	0.315	4	21
			h43	S48	0.481	3	
			h31	S35	0.722	6	
			h38	S42	0.870	8	
d.47.1	**L11**	0.502	H45	D20	0.407	14	18
			H44	D19	0.630	4	
d.55.1	**L22**	0.502	H73	G1	0.296	9	42
			H2	B1	0.389	10	
			H61	E19	0.500	4	
			H50	E6	0.519	7	
			H26	D1	0.556	4	
			H99	J1	0.722	2	
			H3	B2	0.778	4	
			H35	D10	0.944	2	
c.22.1	**L4**	0.507	H27	D2	0.167	10	69
			H46a	D22	0.407	24	
			H19	B19	0.556	10	
			H26	D1	0.556	6	
			H28	D3	0.556	7	
			H22	B21	0.796	12	
e.24.1	**L1**	0.516	H76	G4	0.037	4	4
d.59.1	**L30**	0.516	H41	D17	0.037	17	41
			H38	D14	0.037	14	
			H46	D21	0.556	7	
			H41	D16	0.907	3	
c.12.1	**L15**	0.525	H88	G15	0.426	16	60
			H37	D13	0.444	14	
			H12	B12	0.519	8	

Table 3.2 (cont)

FSF	r-Protein	nd r-protein	rRNA Contact		ndrRNA	Number bases interacting	
			Brimacombe	ErRD			
			H11	B10	0.556	3	
			H36	D12	0.685	12	
			H29	D4	0.926	4	
			H12	B11	0.963	3	
d.12.1	**L23**	0.529	H51	E7	0.907	4	4
a.16.1	**S15**	0.534	h26	S29	0.204	2	29
			h22	S24	0.333	17	
			h20	S23	0.593	5	
			h23a	S26	1.000	5	
j.84.1	**L10**	N/A	H42	D18	0.630	8	8

3.3.2 A factor-mediated second transition precedes the 'big bang' of the protein world.

Consistently, the most ancient FSF domains are universally present in all organisms and with time they are first lost in primordial archaeal lineages and then in eukaryal and bacterial lineages [92, 168, 174]. In turn, the rather late gain of Bacteria-specific, and then, Eukarya-specific and Archaea-specific structures, signal the emergence of superkingdoms. These same patterns were also observed in the diversification of ancient RNA molecules such as tRNA, 5S rRNA and ribonuclease P RNA, with RNA substructures specific to Archaea appearing before substructures specific to other superkingdoms [85, 126, 175]. These patterns revealed the origin of the tripartite world, highlighting three evolutionary epochs [124]: an ancient 'architectural diversification' period (Epoch 1) in which ancient molecules emerged and diversified and proteomes were highly homogeneous, a 'superkingdom specification' period (Epoch 2) in which molecules sorted in emerging organismal lineages, and a late 'organismal diversification' period (Epoch 3) in which molecular lineages diversified and became specific to superkingdoms and notable proteome expansions occurred in Eukarya. In this timeline, reduction of structural repertoires was mostly confined to primordial archaeal lineages at the end of Epoch 1, an observation that is also confirmed in the phylogenomic tree Figure 3.6 A. Remarkably such reductive evolution patterns were also observed in r-protein families [156]. The r-proteins S12 and S17 (nd_P = 0.018) are the oldest and appear at the start of the protein world (Figure. 3.5 A). Similarly, a modern RNP translation apparatus evolved during Epoch 1 concurrently with L3, L2

and L24 (nd_P = 0.05-0.2), long before many other r-proteins, most of which appear together in a narrow time interval (nd_P = 0.40-0.53), and before the rise of superkingdoms and a diversified world.

It is however noteworthy that a recent study of evolutionary mechanisms of domain organization and modularity in the protein world revealed that during early Epoch 1 (nd_P < 0.1) single domain multifunctional proteins dominated [174], domains fused and proteins with fewer

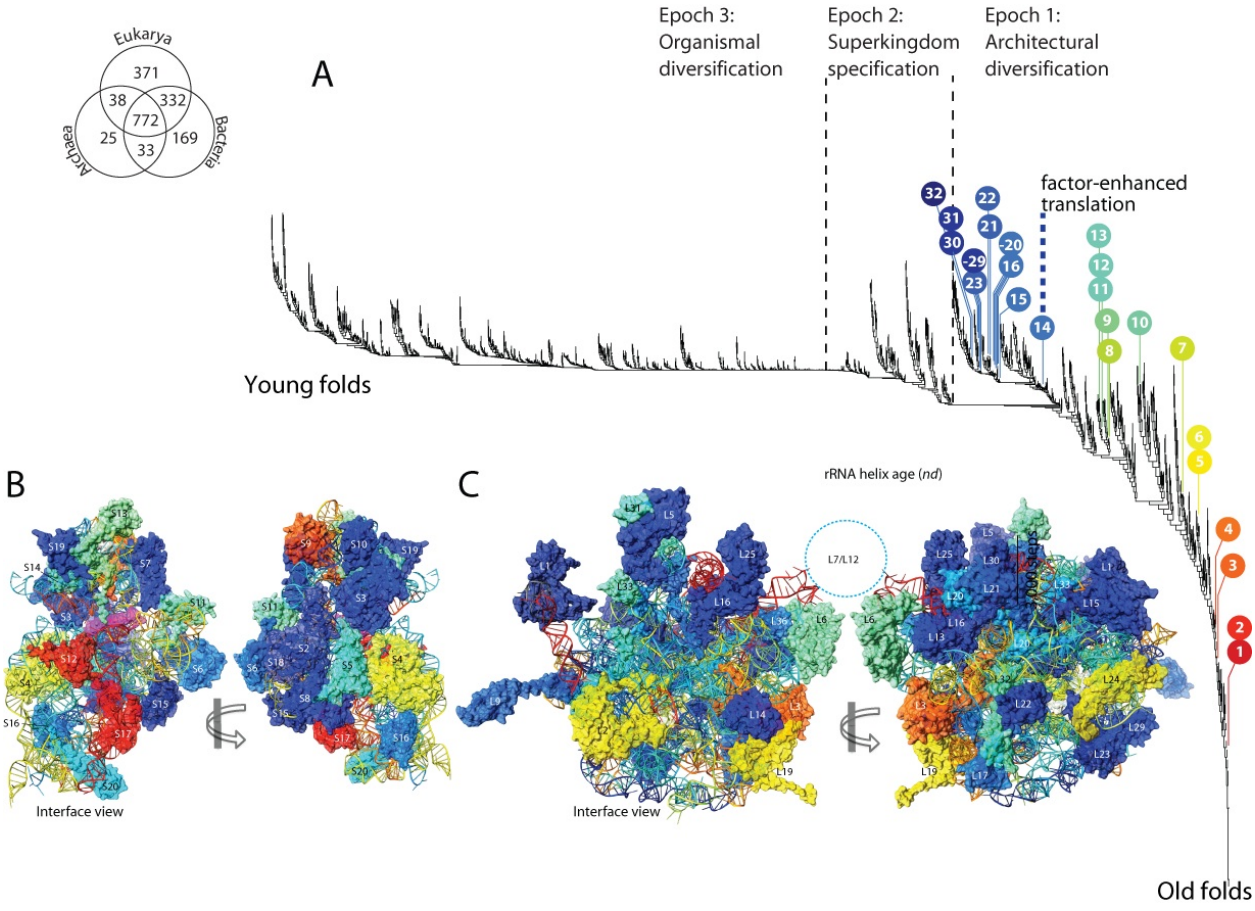


Figure 3.6: Evolution of the r-proteins in the context of protein evolution traced at the FSF level. (A) Backbone of universal tree describing the evolution of 1,730 SCOP FSF domain architectures from 749 genomes (541,383 steps; CI = 0.028, RI = 0.783; g1 = -0.111) shows a relative time line of discovery of protein architectures. r-protein FSFs are colored corresponding to the age as in Figure 3.5 to show their relation to the rest of the proteins world. Other than a few old r-proteins *viz* S12, S17, S9, L2, L3 and L24, most other r-proteins evolved late and in a short span of time, but before the birth of the tripartite world (Epoch 2). The Venn diagram shows occurrence of FSF in the three superkingdoms; FSF common to all life are located at the base of the tree. (B) Evolutionary heat map of SSU r-proteins. (C) Evolutionary heat map of LSU r-proteins. The 3D structures show the relative age of the rRNA segments and the relative age of r-proteins interacting with them. The oldest r-proteins S12 and S17 interact with the oldest helices of the SSU.

functions started to evolve (nd_P = 0.1-0.3), and later on, new domain combinations massively emerged as a result of fusion and fission activities in a 'big bang' (nd_P > 0.6) [174]. However, during nd_P = 0.32-0.40, the fusion of domains and discovery of r-protein FSFs notably ceased. This 'gap' could indicate a fundamental revision of the protein biosynthetic apparatus after which the rate of discovery of new FSF architectures increased drastically. If this drastic improvement could be attributed to a single event, this event would be the enhancement of protein synthesis efficiency by factor-mediated translation driven by GTP hydrolysis. EF-G catalyzed elongation increases protein synthesis more than 50 fold [176]. The GTPase activity of EF-G requires and is strongly stimulated by r-protein L7/L12 [177], which appears at nd_P = 0.42 in our timelines (blue dotted line in Figure. 3.4). It is striking to note that this event corresponds to development of the GTPase associated center of the LSU rRNA and the corresponding age of the L7/L12 protein complex associated with it. The rapid protein diversification seen in the tree of FSFs that occurs in a very defined clade of the tree (Figure 3.6 A) and the change in r-protein-rRNA age congruence (Figure 3.4) can both be explained by the sharp increase in the overall processivity of the ribosome. We regard this as a second major transition in the evolution of the ribosome. We propose that during this revision process proteins refined the structure of the rRNA machinery and increased processivity. The new RNP apparatus was much more efficient than its predecessor. Many experiments that truncate or delete r-proteins resulting in decreased activity of the ribosome confirm our hypothesis [161, 178]

3.4 Chronology of ribosome evolution shows gradual accretion of both RNA and protein domains

The relative age of both rRNA helices and the r-proteins that bind to them were independently determined. Tracing the addition of each of these elements at a given time point during various stages of the evolution will provide a picture if either RNA or protein domains were predominant or if they coevolved as mixed RNP domains. Figure 3.7 shows the accretion of the different rRNA and r-protein components. The complete time of evolution, that is *nd* 0 thorough 1 was divided into 10 time points each increasing by 0.1 *nd* units. Remarkably, the older regions of rRNA bound to older r-proteins in a concerted pattern consistent with coevolution. This is unlike

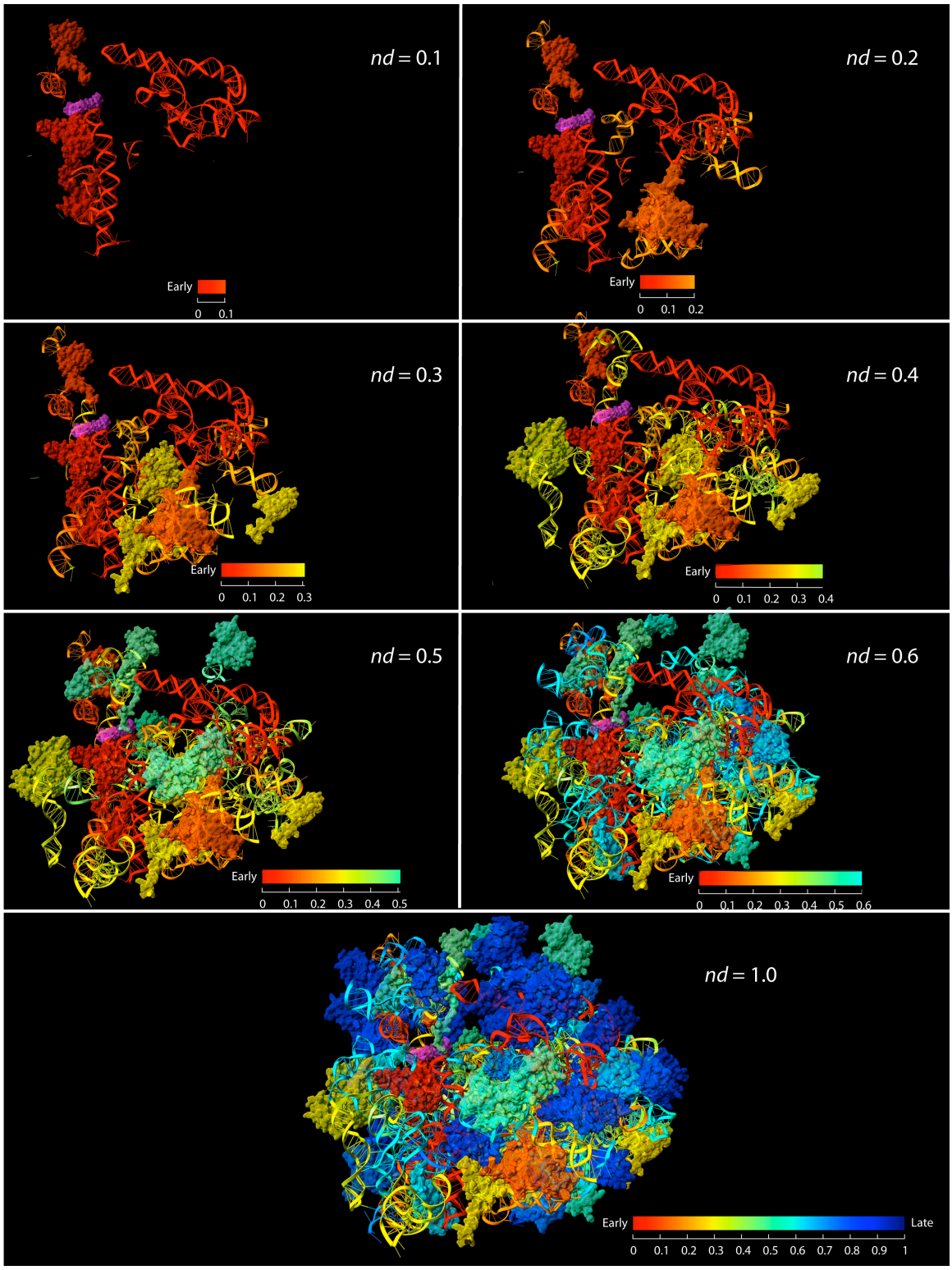


Figure 3.7: A chronological representation of the evolution of the ribosome. At *nd* = 01, many older rRNA helices interact with r-proteins at a very early stage. Interestingly the older proteins that are at the core are also known to perform important functions and their association with the functional center with RNA shows that proteins interacted with rRNA very early during the evolution of ribosome.

previous models that propose a protoribsome solely composed of RNA. As seen in Figure 3.6, after the evolution of the GTPase-associated center and EF-G binding regions contributing the enhanced rates of protein synthesis, there is a rapid increase in the accretion of r-proteins in the ribosome. Almost half of the proteins of the ribosome (blue) appear between *nd* 0.6 and 0.8. coinciding with the development of the GTPase associated center and the L7/L12 mediated factor-enhanced translation.

3.5 Conclusions

Although ribosomes are considered RNA machines [51], it is clear from the data presented here that RNA and proteins have coevolved for longer than usually considered. Given the nature of complexity and cooperativity of the RNA and protein components, RNA-protein interactions are likely to pre-date the modern ribosome. In extant cells most RNA catalysts are always found associated with proteins. Although the present day proteins are synthesized through a ribosomal mechanism, it is possible to retain the structural properties from a time when those proteins were not genetically encoded. Alternatively the structures of rRNA and r-proteins are so 'canalized' that such associations may be possible simply because of an evolutionary 'memory'. However this is only speculative at this point and can be tested.

The assignment of relative importance to either RNA or protein and considering a particular component fundamental depends on mutation or deletion of one protein at a time. Even a small change in some of the proteins reduces the rate of peptide synthesis or decreases the degree of fidelity. Tracing the evolution of both RNA and protein components shows that the relative proportion of protein and RNA remains roughly constant at all stages of evolution of the ribosome. It is remarkable that the relative ages defined by two different approaches at different levels of hierarchy among their structural organizations show such high congruence. Thus the concept of 'fundamental' contribution to translation does not seem to be useful [179]. Neither is the concept of relative contribution useful. The complex as a whole defines the function. Considering the distribution of mutation rates in protein enzymes, it is clear that although disruption of a few residues that define the catalytic center affects the enzyme activity,

the catalytic activity alone does not define the enzyme. The robustness of the catalytic activity depends in the overall stability of the protein, which is a result of 'canalization' of the structures towards increased resilience to perturbation [134]. Ribosomal robustness is in its processivity and accuracy of translating the genetic code. [180-182] Translational robustness thus affects organismal fitness [183, 184]. The genetic code has evolved to be highly optimized and reflects coevolution of tRNA abundance and codon usage [128, 185] and is related to translational accuracy [186] ultimately constrained by aa-tRNA selection and mRNA-tRNA translocation [187].

Although a primitive ribosome composed solely of RNA has been proposed[188, 189], it is unlikely that such a complex RNA machine could have existed . Instead, multiple smaller RNP molecules performing different functions probably integrated in evolution into a much more complex RNP ensemble [127]. Initially, peptide synthesis was very inefficient and had to be non-ribosomal, especially because protein chronologies indicate ribosomal proteins appeared well after metabolic enzymes (Figure. 3.5 A) [124]. Synthesized peptides were short and increased the functional and structural repertoire of RNAs [170]. They were probably unable to fold independently, and maintenance of their conformations required the RNA scaffold [190].

3.6 Materials and Methods

3.6.1 Determining the ancestry of r-proteins

The general scheme applied to study the evolution of RNA structures has also been used to infer evolutionary relationship among protein architectures at the fold (F) and fold super family (FSF) categories in the Structural Classification of Proteins (SCOP) database on the basis of their occurrence and abundance in genomes [92, 166]. The relative age of r-proteins were determined by the ancestries of their FSFs derived from an updated tree published earlier [92]. Each FSF was described by features that numerically characterize their genomic abundance (G). G values were normalized to offset differences in genome size and frequency of each FSF in corresponding proteomes, and log transformed to account for unequal [12] variance. G values were then used as range standardized character states represented as discrete alphanumeric set

with numbers 0-9 and letter A-K and encoded in the NEXUS format. Values were converted into linearly ordered multistate characters using the gap-recoding technique developed for cladistic analysis of morphometric data [191]. Almost all proteins are share structural similarities with other proteins and are related by common ancestry [82]. Phylogenetic and large-scale statistical analyses have shown that convergent evolution of domain architectures is rare [192] and the diversity of protein structures arose from a small set of ancestral peptides by descent with modification. Gene duplications and (domain) combinations give rise to proteins with novel structures and functions [193] and a limited repertoire of frequently combining domains mostly accounts for the diversity and evolvability of protein architectures [193]. Thus it is reasonable to assume that FSF architectures that are successful (selected) and popular in nature (maximum character state) are generally more ancestral (plesiomorphic).

Characters were polarized with the ANCSTATES command. Trees were reconstructed using maximum parsimony as the optimality criterion, and were automatically rooted at the point where the hypothetical ancestor connected to the tree. Since reconstruction of large trees is computationally intensive a combined parsimony ratchet and iterative search strategy was used. Phylogenetic reliability was tested by the non -parametric bootstrap method implemented using 1000 replicates. Relative age (nd) of the FSFs was determined as described earlier in Chapter 2, section 2.8.6. Further, nd of a subset FSFs corresponding to r-proteins were extracted

3.6.2 Evolutionary Heat Maps

To better visualize the relative age of the different elements of the ribosomal ensemble and to understand how the functions associated with these structural elements, secondary structures of rRNA and the 3D structure of the ribosome were painted with colors corresponding to their respective *nd* values. Secondary structure diagrams of *Thermus thermophillus* rRNA corresponding to the crystal structure of the 70S ribosome (PDB id 1GIX and GIY) were obtained from the Noller Lab website at (http://rna.ucsc.edu/rnacenter/ribosome_images.html). A RGB color scale corresponding to the nd values 0-1 with an interval of 0.01 was produced in matplotlib [146] using scripts available at http://matplotlib.sourceforge.net/gallery.html. The

secondary structure models were modified and colored according to the nd values. Helix numbering from the European Ribosomal RNA database was reconciled with the Brimacombe numbering scheme. The crystal structures of *Thermus thermophillus* 70S ribosome (PDB id 2WDK and 2WDL) were also colored according to corresponding nd values of the rRNA helices. However, since the r-protein FSFs were a small subset of a large FSF tree, those nd values were extracted (range nd=0.018-0.534) and normalized to a 0-1 time scale as follows.

> For values A (lowest) to I (highest) to be rescaled between *a* and *i* where A is *a* and I is *i* after transition. For any number n between A and I, inclusive,
>
> let $x = (n - A)/(B - A)$
>
> The rescaled value is
>
> $i*x + a*(1 - x)$
>
> In this case A = 0.018 and I = 0.534. Since $a = 0$ and $i = 1$, the rescaled value is x.

3D evolutionary heat maps were produced using the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco [194-196].

An alternate method of interpolation was used to determine the r-protein *nd* with reference to rRNA helix *nd.* Figure 3.6 shows that the protein interactions follow a linear correspondence with the rRNA helix nd. Starting from the oldest helix and the oldest protein that interacts with the correspondence is maintained until the point of the second transition after which there is rapid burst in the discovery of the new FSFs. Hence the pattern of nd_P and nd_P correspondence is interrupted. To determine the corresponding nd_P for such the newest r-proteins on with a contact to the newest rRNA helix was plotted and linked to the slope of the older protein contacts. The nd_P values were interpolated on the slope as shown in Figure 3.8. In the dataset of universal r-proteins (Table 3.2), most proteins are made up of only one domain. In this case the age of the protein is the age of the domain. However, r-proteins L2, S3, S5, L11 and L10 are made up of two domains. In this case, the second domain added to the protein could be an ancient domain that was co-opted for the new task or it could be a new domain that was recruited to enhance the old function. To distinguish between these two possible scenarios we

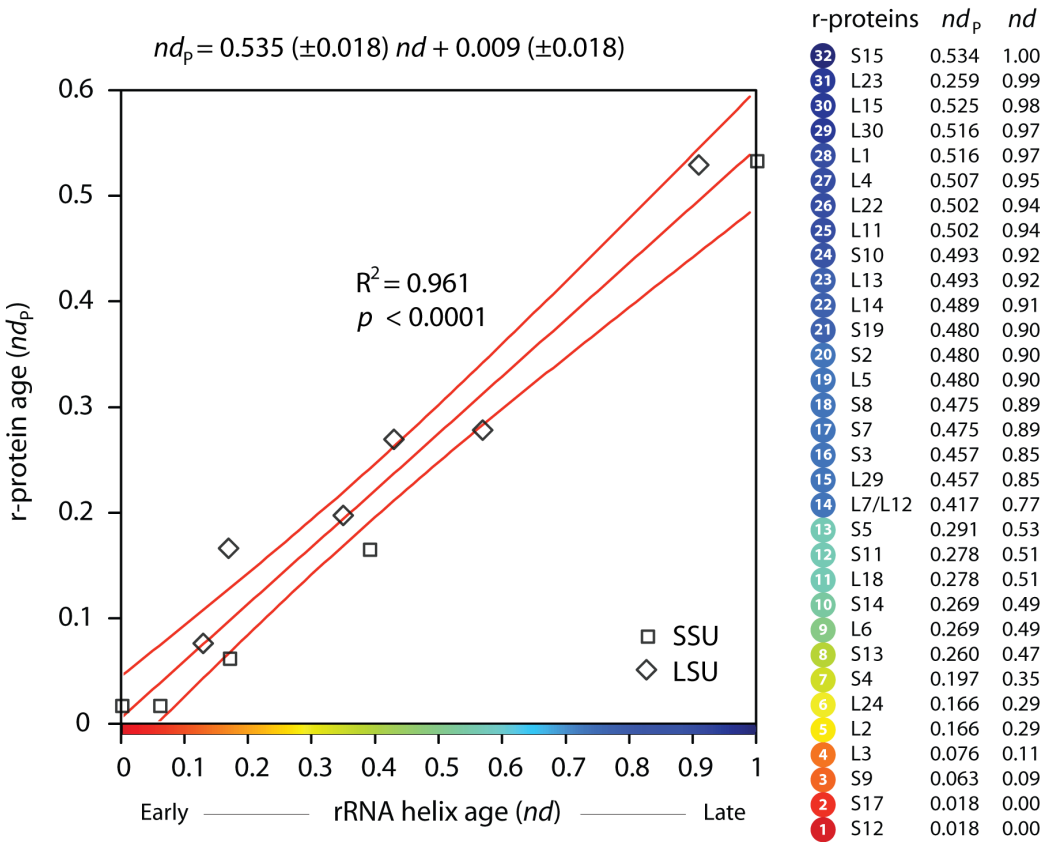


Figure 3.8: method of interpolation was used to determine the r-protein *nd* with reference to rRNA helix *nd*.

examined a tree of domains and domain combinations to determine placement of the two-domain protein in the timelines [174]. For example, if domains **a** and **b** could either fuse as **a-b** or **b-a**. If combination **a-b** were older than **b-a**, then a r-protein with combination **a-b** would be assigned the nd_p of that domain combination. When this information was not available we assigned the age of the newer domain from the FSF tree in Figure 3.6 since the domain fusion in this case could not have occurred until the appearance of the newer protein.

Chapter 4

The search for a primitive replicase

4.1 Introduction

The hypothesis that all life on Earth traces back to a single common ancestor is the central guiding principle in investigating the origins of life on Earth [197]. Although Life could be defined in many different ways [198], efforts are currently focused on deciphering the nature of a 'minimal cell' [199]. This cell is compartmentalized to distinguish itself from others, capable of metabolism to sustain itself, and able to reproduce and evolve to adapt to changing environments. Reproduction is hereditary where encoded genetic information is transferred to successive generations. In all extant cellular life DNA is the genetic material and protein enzymes both replicate the genetic information and perform almost all cellular functions. Genetic information is expressed through an RNA intermediate. Thus knowing which of these molecules came first during the origin of life was a major puzzle until the discovery of catalytic RNAs (ribozymes) in the 1980s [151]. The discovery that RNA molecules can both store genetic information and catalyze chemical reactions was considered a missing link and hypothesized that RNA likely pre-dated DNA and protein. In 1986, 'The RNA world' was thus proposed as a hypothetical stage in the origin of life, although it was conceived much earlier [188, 189, 200, 201].

The RNA world is interpreted in many ways but can be broadly categorized into two types of hypotheses [202], One school of thought proposes primitive life forms that used only RNA to genetically encode biological catalysis were intermediates, which could have been preceded by a variety of catalytic systems other than RNA.

The second school is rather too 'RNA-centric' and much less supported. It proposes that a primitive life form that solely used RNA to genetically encode biological catalysis was the first form of life on earth, the first chemical system on earth to support Darwinian evolution.

The difference between the two schools is that the second requires that the RNA world emerged spontaneously from a prebiotic inorganic world, while the fist one proposes that RNA, among the many different organic molecules that could have followed prebiotic chemistry, is the most recent. In general all forms of the RNA World hypotheses have the following basic assumptions [203]

- At an early stage in the origin of life, RNA was the most suitable molecule capable of replication and hence provided for genetic continuity required for Darwinian evolution.
- Complementary (Watson-Crick) base-pairing was the basis of replication, and
- Genetically encoded proteins were not catalytically active.

Critics of the RNA world however contend that the limited range of catalytic reactions and very low efficiency of RNA compared to protein catalysts and have proposed an alternate proteins-first origin of life [204], the 'Protein World' [179], detailing scenarios of how peptide based life could have started and how such systems could give rise to RNA/DNA based genetic systems.

4.2 Top-down and bottom-up approaches to deduce the 'minimal cell'

There are generally two methods that have been used to understand and estimate the cellular componentry of a 'minimal cell'. The top-down approach tries to systematically simplify existing unicellular organisms [199]. At present it is by genetic methods that delete multiple genes in a stepwise process and in an effort to arrive at a minimal genome. These efforts could benefit from phylogenomic methods that reconstruct the genome of a hypothetical ancestor [205]. For instance the protein architecture tree could be used to root a universal tree and to

determine the minimal proteome. In addition, reconstructing a tree to deduce the minimal regulatory genome can complement the approach. However the search for a universal cell would be a futile effort since organisms have diverged far away from a common ancestor and hence models at different levels of biological organization have to be tried [199]. Although phylogenetic methods have not been used yet to resurrect an universal cellular ancestor, it has been successfully employed to infer ancient genomes [205, 206] and to resurrect individual genes and proteins to understand various aspects of their evolution. Resurrected ancient Ef-Tu was found to be functional and was used to infer ancient environments and lifestyles [207] and resurrected short-wave opsins were used to infer the evolution of novel functions related to color vision [208]. These are only a few examples among many.

The bottom-up approach aims to assemble chemical systems that are capable of self-sustenance. Progress in this direction has been made recently in creating micelles capable of reproduction [209]. Most of the efforts with this approach are motivated by the RNA world hypothesis and one of the important pursuits at present is to select an in vitro evolved ribozyme capable of replication [210-212]. A variety of ribozymes capable of a multitude of chemical reactions have been selected after thousands of rounds of iterative mutation and selection to mimic the natural process of evolution [49]

4.3 Evolution of coded protein synthesis

The origin of the genetic code and translation is considered to be a 'major transition' that transformed a primitive life to a modern protein based life. Considering an RNA World, it meant using genetic information in a radically new way and allowing for a division of labor between nucleic acids and proteins that serve as genes and enzymes respectively [213]. Whether life started with RNA or proteins, explaining the origin of the collinear relationship of nucleic acids and the proteins corresponding to the triplet genetic code has been a very difficult problem. "*The origin of protein synthesis is a notoriously difficult problem*" was the now famous remark from Crick summarizing the complexity of protein biosynthesis [214]. Earlier attempts were highly speculative. The origin of translation was reduced to the origin of the genetic code and separated

from evolution of the translation apparatus [116]. The genetic code was called a "frozen accident". The assumption that the evolution of the genetic code and protein synthesis go hand-in-hand has been criticized [215]. It is considered incomplete and insufficient to understand the origin of coded protein synthesis [116]. Moreover, it has been proposed that such ideas have hindered the progress in the field. Furthermore it has been suggested that genetic code-centric approaches based on Watson-Crick base paring have misguided the understanding of evolution of 'gene expression'. "*A satisfactory level of understanding of the gene should provide a unifying account of replication and expression as two sides of the same coin. The genetic code is merely the linkage between these two facets*." [216]

Modern translation is incredibly complex in terms of processes and the number of molecules involved. Such a complex system had to have evolved in many stages where each preceding stage was less complex [116]. According to the principle of continuity the roots of such a complex system can be traced to preexisting function(s) [50, 217]. That is the potential for a complex coordinated process must have preexisted even if in a rudimentary form. Thus the genetic code is predicted to have evolved before the evolution of translation and for a different reason (function). Many models have proposed that the origins of translation is linked to replication. These are highly speculative. Nevertheless, proto-ribosomes are proposed to be primitive replication apparatuses [50, 217-220]. Particularly interesting are models that couple origin and coevolution of the genetic code and RNA replication facilitated by proto-ribosomes [221-223]. Perhaps the most comprehensive and well-conceived hypothesis that addresses every aspect of ribosomal function to its hypothetical predecessor RNA polymerase is the triplicase hypothesis [217]. The model is as follows. In an RNA world before the advent of coded protein synthesis, the proto-ribosome could have been a RNA-based RNA-replicase with reasonably high fidelity. Some features of the proto-ribosome are as follows

- To perform complex functions it was relatively large ~4500 bases whose length did not increase further after proteins were incorporated.
- The proto-ribosome could recognize and bind to single stranded RNA (mRNA equivalent)
- tRNA precursors that bring in nucleotides to be incorporated in the newly synthesized RNA, 3 at a time to increase binding and specificity of the reaction by allowing a longer time for

the reaction. This is more accurate as compared to a single nucleotide addition. Assumption supported by slow rate of in vitro evolved RNA polymerases.

- A ratchet mechanism that moves the single stranded RNA in steps of three. Supporting experiments show that protein reduced SSU can efficiently translocate mRNA in ribosome [161].

Furthermore, they point out that the evidence that tRNA or its precursor was a donor of nucleotide is from the tRNA processing in extant organisms, a tRNA intron is cleaved from a pre-tRNA at the anticodon site to form a mature tRNA anticodon stem loop. Since it adds three nucleotides at a time and replicates, it is called a 'triplicase', which is both a ligase and polymerase. Finally, according to the model this proto-ribosome itself evolved gradually where the proto-SSU could be the RNA-polymerase and the proto-LSU initially formed the ratchet mechanism. Presumably basic amino acid tagged tRNAs could stabilize the interaction of negatively charged nucleic acids. The genetic code was not fully developed at this stage.

The RNA triplicase theory is attractive because it provides for a high fidelity replicase/polymerase, an origin for a triplet code, and an origin for the ribosome. It predicts that rRNA/mRNA/tRNA interactions are ancient and pre-date proteins.

Apart from the ribosome there are no examples of natural RNP polymerases. Ribosomes and DNA/RNA polymerases use similar strategies of minor-groove recognition to maintain fidelity. Fidelity permits an error-prone primitive self-replicating system to evolve into a complex system [51]. Interestingly, fidelity and processivity are tightly linked in ribosomes [224]. The absence of natural RNP replication enzymes represents a gap in evolutionary continuity and precludes the possibility of obtaining a natural phylogeny of RNP and protein polymerases. However, *in vitro* selected ribozymes substitute as doppelgängers for supposedly extinct molecules and provide means to test the likelihood of their existence [49] . Recent crystal structures of two such ribozymes involved in ligation and polymerization of RNAs have helped understand the reaction mechanisms [210, 212]. Moreover, both natural and artificial functional RNAs share universal evolved sequence features that inherently define conformational order [225]. These features arise due to intrinsic properties dictating RNA self-organization and

not from selection [47, 226]. Hence detection of any remote homology between rRNA substructures and replicase/polymerase dopplegängers would support the role of proto-ribosomes in replication. Implicit in this reasoning is the assumption that evolutionary history from a time when the proto-ribosome is hypothesized to be a RNA polymerase is still preserved in the rRNA structure of extant ribosomes.

To test these hypotheses, we used a RNA structure comparision tool find evidence of structural homology between the older regions of the ribosome (processivity core) and *in vitro* evolved ribozymes such as the L1 ligase and RNA polymerase. Natural ribozymes such as RNase P was also studied. Finally, the hypothesis that tRNA was the precursor of modern rRNAs [68] was tested with this analyses.

4.4 Results and Discussion

Many RNA secondary structure comparison algorithms have been developed recently [227-231]. However none of them are specifically developed to compare or infer relationships between divergent structures. All of them have stemmed from efforts to predict and identify non-coding RNAs in genome sequences exploiting the higher degree of structure conservation in such gene families with highly divergent sequences [232]. Hence although these algorithms were developed to exploit RNA structure conservation features, most use scoring schemes and similarity measures that rely on the sequence rather than the structure. After a survey of various methods with predefined, custom mutated sequence-structure pairs, RNAforester [233] was found to be most useful, particularly to detect local similarities in RNA secondary structures. RNAforester comes with an option of a scoring scheme that is purely on the structures being compared. In addition it was found to be most robust in detecting distantly similar secondary structures as well as closely related ones.

RNAforester is essentially an equivalent of the Smith-Waterman (S-W) algorithm [234] applicable to RNA structures. However, unlike the S-W algorithm, there are two major differences in the alignment and scoring scheme. First, in the alignment scheme, a secondary structure is first converted to a tree (called forest representation) where the internal nodes

represent paired bases and the terminal nodes represent unpaired bases in the structure. Instead of the conventional string alignment a more robust tree alignment is used to facilitate differential scoring of paired and unpaired bases in the structure. Since paired regions define the structure, a higher score is assigned to a base-pair match or mismatch and a much lower score for an unpaired base. Second, scoring is dependent on edit distance instead of alignment distances and sequence contribution to the score is negligible. As opposed to alignment distance in conventional sequence alignment algorithms such as those implemented in BLAST, to compute an edit distance between two strings being compared, one string is "edited" into another string by a sequence of edit operations, such as deletion, insertion or substitution. The weights associated with the edit operations sum up to an overall score. The edit sequence giving the minimal score defines the edit distance of the two strings.

4.4.1 The ribosomal core and ribozyme doppelgangers

To detect this remote homology between older regions of rRNA and doppelgangers we used RNA secondary structure similarity searches. Pairwise structural alignments of *in vitro* engineered doppelgangers (experimental mimics of extinct RNA) [49] and all rRNA substructures were used to probe the function of the ancient scaffold and of the proto-ribosome. Hypothetical ancestral SSU and LSU rRNA sequence and structures reconstructed directly from our trees were aligned to L1 RNA ligase ribozymes (RL) [210], RNA-polymerase ribozymes (RP) [212], aminoacyl-tRNA synthetase ribozymes (AARS) [235] and natural functional RNAs RNase P and tRNA using an advanced RNA structure alignment software RNAforester [233].

Only those molecules whose crystal structures are available and hence confirmed secondary structures were used. The L1 Ligase is an in vitro evolved ribozyme selected from a pool of synthetic random sequences. The ribozyme catalyzes the template dependant 5'-3' phospho-diester bond synthesis required to ligate nucleic acid monomers or oligo nucleotides [210]. The RNA-polymerase ribozyme is also a sequence isolated from a random pool and selected for improved activity by multiple rounds of mutation and selection. Given a template and a primer, the ribozyme extends the primer accurately [212].

The core of the ribosome composed of the oldest rRNA helices is hypothesized to be part of a primitive RNA-polymerase [217]. Assuming that history is still preserved in the structure, the RL and RP ribozymes should have the highest degree of similarity to the core helices compared

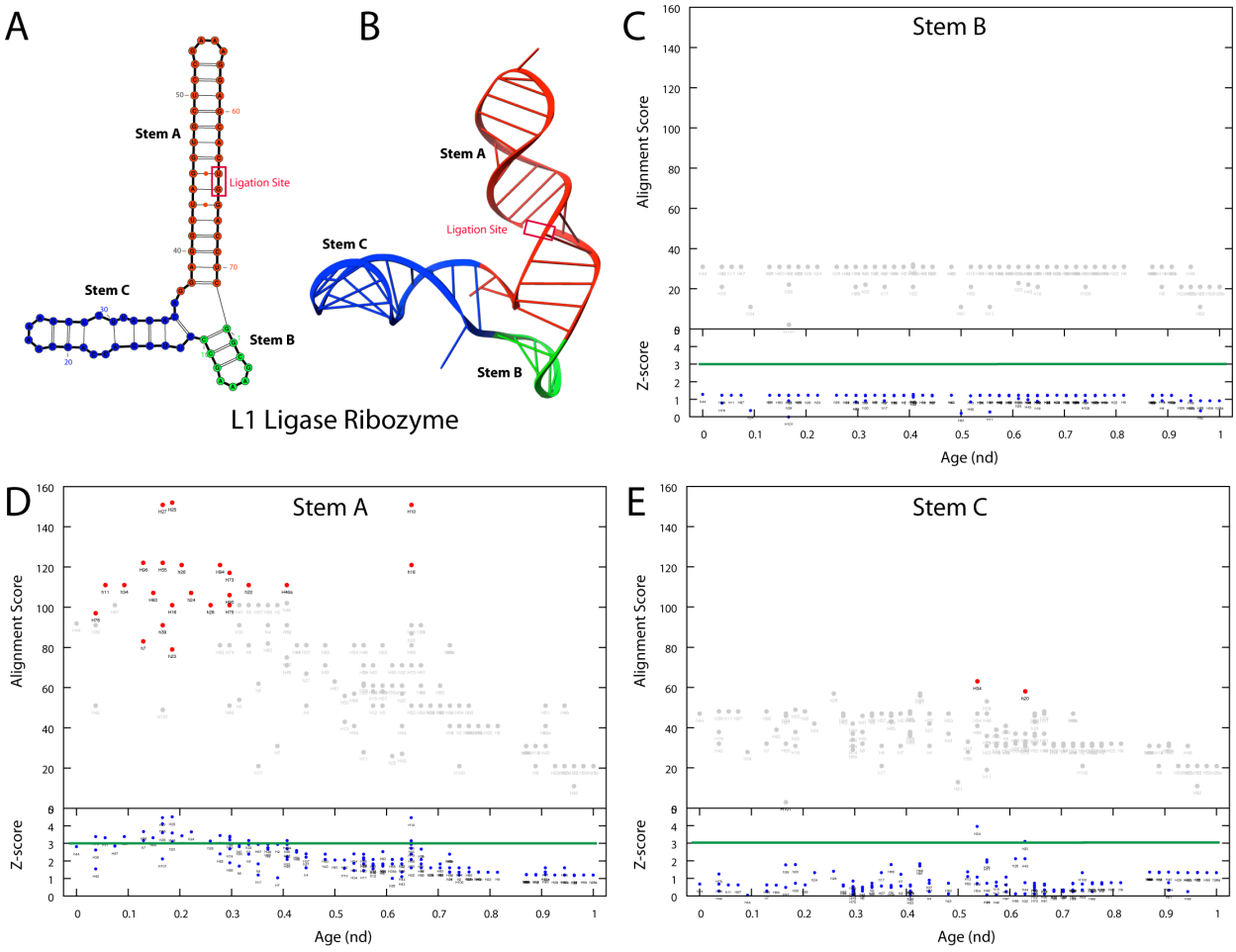


Figure 4.1: Remote homology of hypothetical rRNA ancestor helices and in vitro evolved L1 Ligase. rRNA substructures defining the processivity center have a similar structure and probably similar function (A) secondary structure and (B) tertiary structure of the L1 Ligase [210] is composed a long helical Stem A connected to two shorter helices in a three-way junction. The catalytic ligation site is boxed. (D) RNAforester scores for alignment (y-axis) of the catalytic Stem A to the different rRNA helices arranged by their Age (nd) (x-axis). Significant matches are in red. Lower panel shows a test of statistical significance of the alignment scores. Structures derived from 1000 random sequences preserving the dinucleotide frequency of Stem A sequences were aligned to each of the rRNA helices. Z-scores are plotted for each corresponding rRNA helix. The line indicates a Z-score threshold of 3, corresponding to 0.01% probability that the alignment is by chance. (C) and (E) Alignment scores of Stem B and C. The scores are below the threshold and not a significant match.

to other ribozymes or RNA structures. As the rRNA structure was decomposed to determine the relative ancestry, the helical elements of the ribozymes were also decomposed into

their corresponding helical elements. To minimize any sequence bias and reduce the number of possible alignments required to test structural similarity, hypothetical ancestral SSU and LSU rRNA sequences were reconstructed using maximum likelihood methods implemented in PAUP. Similarly the structure of the hypothetical ancestor was reconstructed. The sequences were aligned with starting sequence alignment. The lengths of the helical elements were then manually corrected to correspond to the hypothetical ancestor structure. The reconstruction procedure is explained in Figure 4.7. The resulting helices from rRNA and ribozymes were aligned pair-wise to determine the degree of similarity.

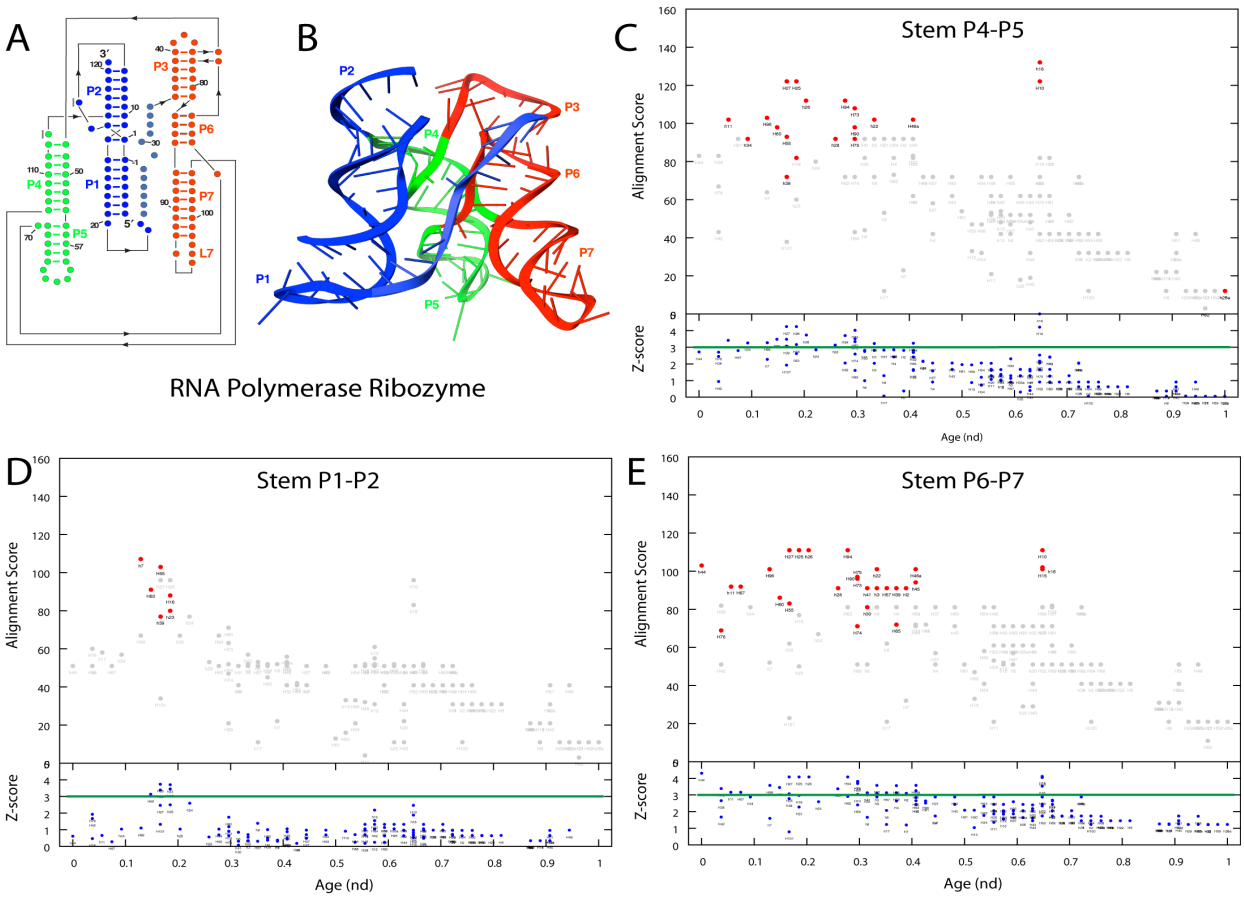


Figure 4.2: Remote homology of hypothetical rRNA ancestor helices and *in vitro* evolved RNA polymerase. (A) and (B) show the secondary and tertiary structures. The catalytic core is at the junction of all three helices Further details as in Figure 4.1

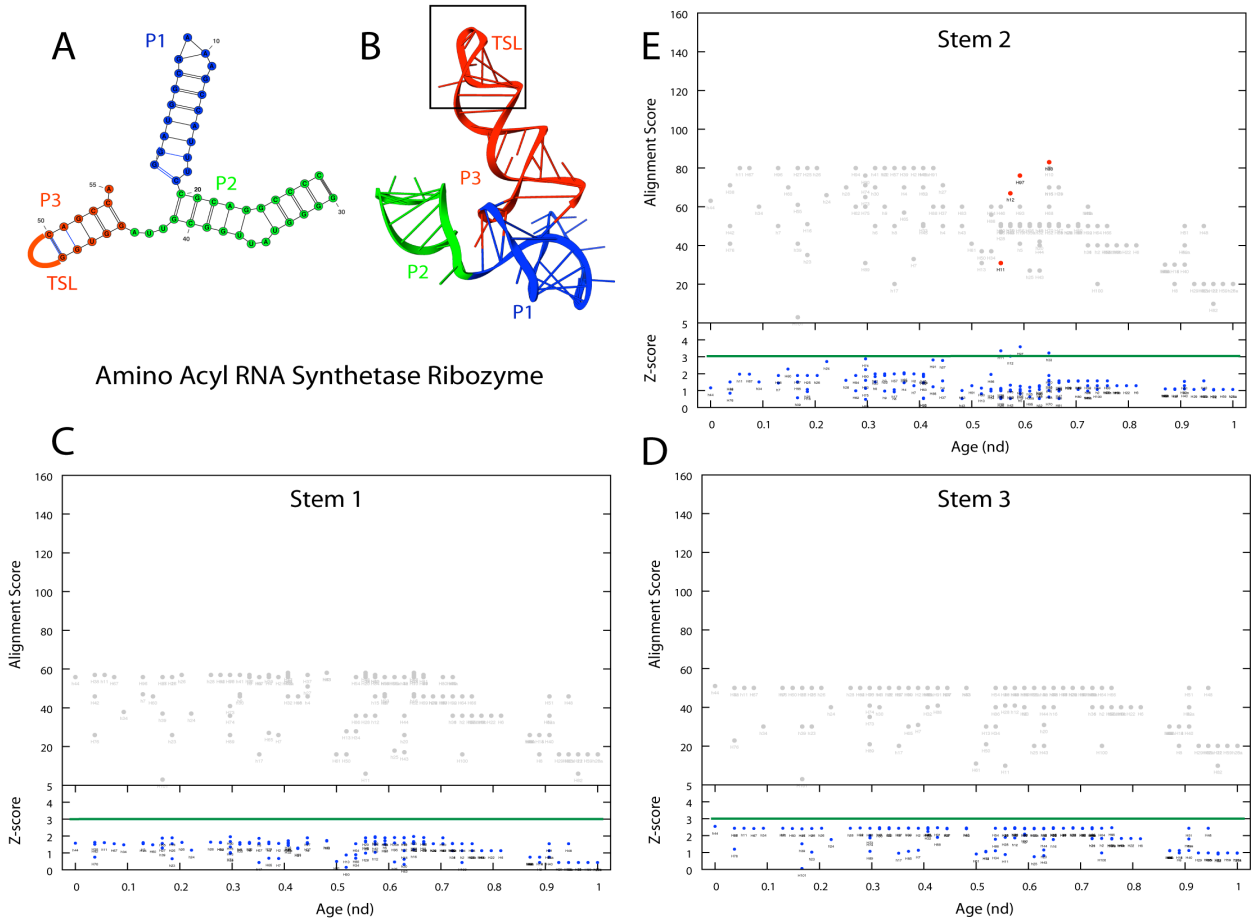


Figure 4.3: Remote homology of hypothetical rRNA ancestor helices and in vitro evolved Amino acyl RNA synthetase. A and B show the secondary and tertiary structures. Further details as in Figure 4.1

Statistically significant homology was detected between the primordial-core rRNA helices (nd ~ 0.3) with the catalytic helices of the RNA ligase, RNA polymerase and RNaseP. This result indicates that the proto-ribosome could have been a replication apparatus. In addition recent biochemical experiments have shown that the catalytic center is composed of two layers of highly conserved unpaired bases that are important for the catalytic activity [236]. However mutational analyses showed that these conserved bases are required for tRNA hydrolysis but peptide bond synthesis is unaffected. This indicates the plausible role of tRNA and the proto-ribosome in a primitive replication process. Our results in combination with this recent experimental evidence provide strong evidence that the catalytic core of the ribosome could once have been part of primitive replication machinery. Since structural components of a proto-

ribosome involved in tRNA interaction, mRNA interaction and intersubunit interactions are older than others, our results support RNA triplicase theory and models of tRNA origins in RNA replication [88, 237].

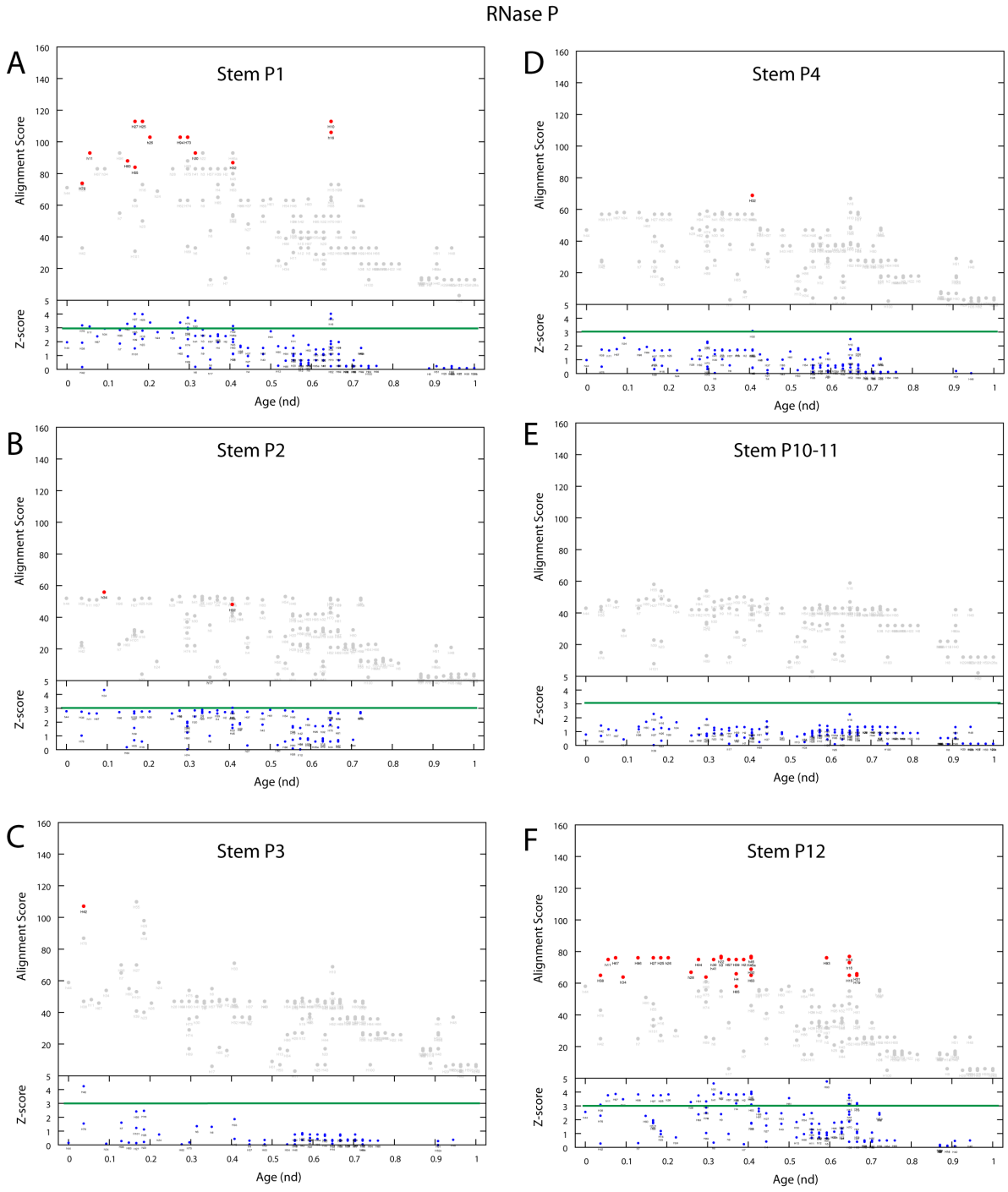


Figure 4.4: Remote homology of hypothetical rRNA ancestor helices and RNase P segments. Helices that were known to be the oldest in RNase P from a previous study [126] were used for the structure homology search. Stema P1 – P4 are in the catalytic domain and stema P10-11 and P12 are in the specificity domain. Further details as in Figure 4.1

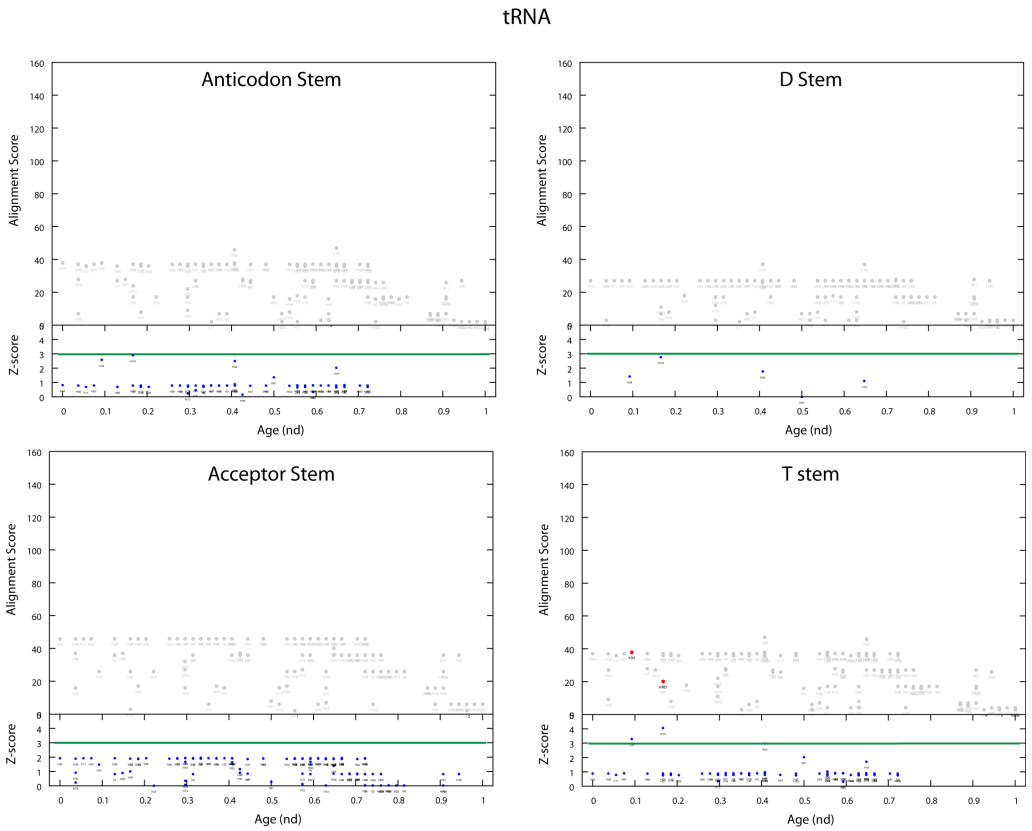


Figure 4.5: Remote homology of hypothetical rRNA ancestor helices and tRNA stems. Further details as in Figure 4.1

4.4.2 Remote homology of catalytic domains to the primitive core of rRNA

Among the many different segments of the natural and synthetic ribozymes analyzed for homology between rRNA and ribozymes, the best match was found be the catalytic domain of the L1 RNA ligase. Matches were significant for the catalytic core of the RNA polymerase ribozyme and also with the catalytic domain of RNase P. Catalytically active domains of both natural and synthesized ribozymes, generally at the heart of the structure of the ribozymes show a remote homology but not other segments of the ribozymes. The comparison of the homologies of different catalytic and non-catalytic segments is shown in Figure 4.6. Interestingly most of the matches are with older rRNA helices (*nd* = 0.0 – 0.35) and predominantly with the catalytic segments of Ligase and Polymerase (read and blue bars in Figure 4.6). However, both catalytic and non-catalytic segments of RNase P showed a match, they are indistinguishable. Hence it was not possible to assign a role to such elements if they were part of a proto-ribosome.

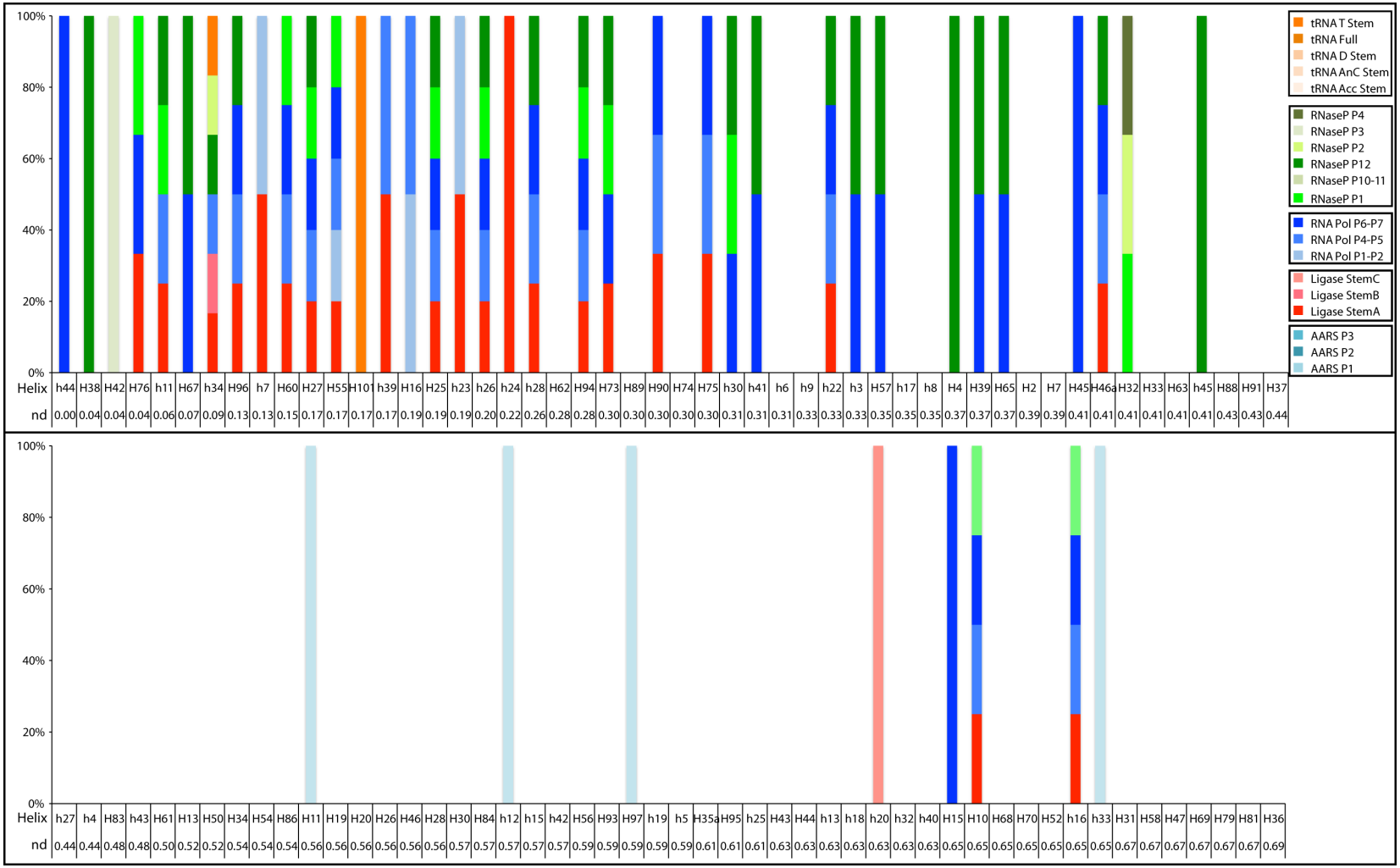


Figure 4.6: Comparison of the high scoring matches of doppelgänger segments with ancestral rRNA helices. Each group of helical segments of respective ribozymes are grouped together in the legend and have different shades of the same color. The red and blue bars are the catalytic elements of Ligase and Polymerase ribozymes. They e have the maximum number of significant matches to the processivity core of the ribosome.

4.4.3 OB-fold proteins and ribosomal origins

In addition to the arguments put forth above, the involvement of the oldest r-proteins S12, S17 of SSU and L3 and L2 of LSU in different aspects of ribosomal processivity and extra-ribosomal functions related to replication further support the functional co-option hypothesis. For instance S12, the oldest r-protein is involved in mRNA movement, tRNA translocation and forms the signal relay for subunit communication effecting the recognition of the correct tRNA to the EF-Tu during decoding [238]. S17 is among the first proteins to stabilize 16S rRNA conformations nucleating the SSU assembly process [65] . Likewise L3 is important to maintain conformation of PTC and is an allosteric switch modulating the binding of the elongation factors [239] and L2 in addition to being important for subunit association [240] associates with RNA polymerase [241] to modulate transcription. Interestingly the oldest r-proteins are part of a family- the OB-fold and the related SH3 domain proteins. Translation initiation factors, tRNA binding proteins including aminoacyl RNA synthetases and DNA binding proteins like the T7 DNA ligase are all part of the family [242]. Interestingly the root of the tree inferred is a common ancestor of RNA binding and DNA binding proteins. Thus, we conclude that r-proteins and homologs were part of primitive replication machinery, which diversified, and developed to completely replace the proto-ribosome based replication apparatus while it was co-opted for translation. Alternatively, a common template directed process could have given rise to both translation and replication

4.5 Conclusions

Clearly the similarity of rRNA helices to L1 ligase and RNA-polymerase detected in alignments indicates that the oldest rRNA helices at the processivity core may have been part of a primitive RNA dependent RNA-polymerase. In addition many aspects of the various steps during ribosome function have mechanistic similarities to DNA/RNA polymerases as follows. Kinetic studies show that codon-anticodon base paring initiates translation elongation and accelerates the induced-fit substrate selection. Other template directed enzymes like RNA and DNA polymerases use similar mechanisms [129]. In addition the mechanism of template recognition conferring fidelity is similar in the ribosome and DNA/RNA polymerases [51]. Moreover, the

movement of tRNA in the 30S subunit limits the overall rate of translocation [243]. Thus we contend that some degree of accuracy of tRNA selection was necessary for template-directed protein synthesis and justify that our model of the origins and evolution of the modern ribosome is centered on tRNA and SSU structural components. Furthermore, accuracy of selection, the rate of selection and the direction of the tRNA-mRNA translocation is greatly enhanced by the r-proteins and translation factors [149, 243] and supports our interpretation of a very early RNA-protein cooperativity. Finally, evidence for a proto-ribosome-tRNA centered replication apparatus can be found in many aspects of mRNA-tRNA translocation during translation. The accuracy of mRNA-tRNA translocation requires an aa-tRNA in the P-site [244] and the SSU E-site is crucial in maintaining the reading frame [245]. The secondary structure and tertiary interactions rRNA have evolved for specific intersubunit communication that follows the deacylation of the A-tRNA during translocation [246]. These aspects are consistent with the 'triplicase' model proposed for primitive replication apparatus that could potentially be co-opted for translation [50]. However, a RNP complex with peptides from non-coded origins is favored. A large complex made entirely of RNA as proposed by the triplicase theory would be unstable.

4.6 Materials and Methods

4.6.1 Ancestral sequence, structure reconstruction and structure alignments

To detect remote homology between the structural elements of rRNA and ribozyme doppelgangers we used a structure alignment tool RNAforester. RNAforester is designed for pair-wise and multiple RNA secondary structure alignments [233]. It is capable of detecting similar structural motifs solely based on conserved structure, independent of their sequence conservation and position. Although scoring is solely based on structural similarity, sequence information can be used to improve the alignments.

In order to simplify the structure comparison exercise and to minimize effects of sequence variation in the large number of rRNA sequences used in the study, a hypothetical rRNA ancestor sequence and structure was reconstructed using the maximum likelihood methods implemented in PAUP (Figure 4.7). We reasoned that a reconstructed model is better than a consensus model. Most parsimonious species trees of the 102 SSU and LSU rRNA structures

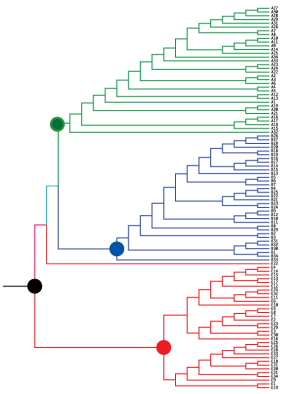
(listed in Table 6.2) were reconstructed using the 'shape' characters. The corresponding DCSE sequence alignments were then converted to FASTA and NEXUS format with SeqVerter for use with PAUP*. Ancestral sequence for the hypothetical ancestor at the root of the tree was determined by reconstructing character states of all internal nodes with the 'describe trees' function and maximum likelihood methods in PAUP. The best-fit model of nucleotide substitution (GTR+I+G) was selected using jModeltest v 0.1.1[247]. The reconstructed sequences were then manually aligned to the DCSE alignment to obtain an alignment based on the secondary structure of the rRNA. The structure was then manually encoded into the Vienna format for use with RNAforester. Similar reconstructions were obtained for tRNA (from 571 sequences) and RNase P (from 133 sequences). The reconstructed sequence and structure was further decomposed into individual helices corresponding to the helical structural elements defined for tree reconstruction in Figure 2.5. The secondary structures of the ribozyme doppelgängers were similarly decomposed into individual helices as defined by the crystal structures. Pair-wise local alignments were performed with each rRNA helix and the helices of the ribozymes doppelgangers. Alignment scores of rRNA helices of different relative ages were compared to determine which ones had the best match.

4.6.2 Test for statistical significance

To determine the statistical significance of these alignments a background model of the structures derived from randomized sequences of the doppelgangers were also aligned to the rRNA helices. A total of 1,000 randomized sequences that preserve the dinucleotide frequency and sequence composition were generated using tools developed by Clote et al as described [248]. Secondary structures of the randomized sequences were determined by RNAfold [249] from the Vienna RNA Package **v**1.8.4. The obtained structures were aligned with the reconstructed rRNA helices for local similarity using RNAforester. Statistically significant alignments were determined based on z-scores. A threshold Z-score of 3.0 was used to determine if the similarity measure bases on RNAforester alignment scores are statistically significant. This threshold estimates that the probability that the alignments were obtained by chance is 0.01. Z-scores are quite commonly used as a measure of statistical significance of alignments when expectation value (e-value) statistics are not available [250].

Process of Sequence and Structure Reconstruction to obtain a Hypothetical Ancestor rRNA

(1) Start with Universal tree of Molecules from Shape Characters (102 sequences)



(2) Convert DCSE alignment to FASTA alignment

```
A U U[C C - G G U]U - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A U - - U G C U]- -
G G A[G A - G U U]U - - - - - G A[U - C C U]-[G - G C U C]- A[G G G C U{G}A A - - C G C U]- -
U A C[C U - G G U]U - - - - - G A[U - C C U]-[G - C C A G]U U[A G U{- C}A U A - - U G C U]- -
- - - - - - 1 - - - - - - - - - - - - - 2 - - - - - - - 1'- - - - - - - - - -3- - - - - - - - - -
```

AUUCC-GGUU-----GAU-CCU-G-CCGG-AGGC--CAU--UGCU

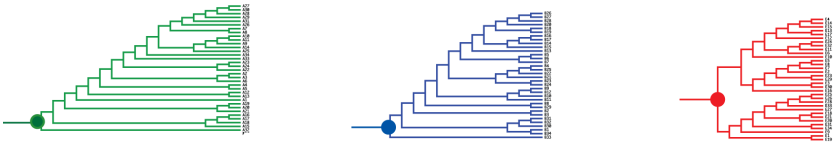
GGAGA-GUUU-----GAU-CCU-G-GCUC-AGGGCUGAA--CGCU

UACCU-GGUU-----GAU-CCU-G-CCAGUUAGU-CAUA--UGCU

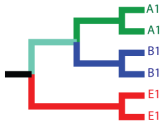
Statistical selection of best-fit model of nucleotide substitution using jModelTest

To reconstruct the Universl Ancestor sequence ●, ideally an alignment of 102 sequences in (1) should be used. However, due to significant sequence divergence between Archaea (A), Bacteria (B) and Eukarya (E) a good alignment was not possible. Hence the ancestor of each superkingdom was first reconstructed from corresponding alignments and subtrees.

(3) Reconstruct ancestor sequence of Archaea ●, Bacteria ● and Eukarya ● using Maximum Likelihood methods in PAUP* and convert it back to DCSE format. Then align structural elements manually.



(4) Use these reconstructed sequences to obtain the Universal Ancestor ●



Since PAUP* requires more than 3 taxa to build trees, duplicates of each A, B and E ancestors were used.

(5) Reconstruct the length of Stems, Bulges and Loops for ● and use this determine the size of stems, bulges and loops of the structure of hypothetical ancestor. Correct the sequence and structure accordingly.

Figure 4.7: A flow chart of the methods and data used to reconstruct the ancestral rRNA sequences from a structure based, rooted tree, reconstructed from sequences from species listed in table 6.2

Chapter 5
Synthesis: Origins and Evolution of the Ribosome

5.1 Revisiting the thermodynamic theory of life.

Theodosius Dobzhansky's assertion "*Nothing in biology makes sense except in the light of evolution*" [251] is perhaps the most cited quotation with reference to the study of evolution besides Charles Darwin's *Origin of Species* [4]. Currently, "evolution" has simply come to mean, "change" [252]. Biology textbooks define "biological evolution" as *change in the properties of groups of organisms over the course of generations*.

In their work "*Life as a manifestation of the second law of thermodynamics"*, Schneider and Kay explore thermodynamic evolution and extend the theory widely from primitive physical systems to complex living systems [45]. They maintain that similar processes, which are manifestations of the second law of thermodynamics, drive evolution of all systems. Their theory was originally developed to understand and explain development of ecosystems but it can be generally applied to physical, chemical and biological systems. The reformulated second law states that when systems moved away from equilibrium, they will use all means to resist externally applied gradients and in the process highly ordered complex systems emerge. They grow and develop at the expense of increasing disorder in higher levels in the system's hierarchy. They contend that as ecosystems grow and develop, they should increase their total dissipation, develop more complex structures with more energy flow, increase their cycling activity, develop greater diversity and generate more hierarchical levels, all to dissipate energy.

Bénard cells are among the best known examples of self-organizing systems observable in real time [253, 254]. A working solution, usually silica oil is placed between a heated plate

and a cooling plate. As the heat is increased gradually, at a certain threshold, there is transition into highly organized convection cells spontaneously. The experiment shows how systems dissipate energy by increasing order.

What is stated above is to show the vast differences in perspectives and theories about evolution. Classical evolutionists were concerned with macroevolution, neo-Darwinian synthesis included classical genetics into evolutionary theories, and the advent of molecular biology influenced a gene-centric view of evolution. Woese and Goldenfeld argue that in spite of the perceived success, reductionist views and dogmatic effects have hindered progress in study of evolution during the past century [216]. They contend that molecular biology's narrow view ignored the process of evolution by over emphasizing narrow mechanistic views. Succinctly, they emphasize that evolution was considered an enhancement when evolution should have been the essence of processes. In particular they describe how the evolution of translation was not well understood because of the narrow perspectives. In trying to explain the parts of the translation system, a view of emergent phenomenon and organization were overlooked. Translation, they consider is an emergence of an incredibly complex mechanism that extracts information from sequences of a kind of macromolecule to express information in the sequence of a different kind of macromolecule. This process if encoding information can continue further to successively higher levels of organization, ultimately giving rise to cells.

5.2 The nature of a common ancestor inferred from tracing the evolution of structure.

In recent times evidence for a complex ancestor has been well supported. Philippe and Forterre's proposal that life evolved through alternating phases of simplification and complexification is finding support [124, 130, 255]. Trees reconstructed from RNA structure and census of protein structures support an advanced and eukaryote-like ancestor. For example, a global phylogeny determined based on protein domain combinations rooted the tree near unicellular eukaryotes unlike the bacterial rooting [256]. In addition large-scale comparative genomics studies have

shown that in many instances gene loss and simplification are common. It has disproved both the bacterial rooting and the hypothesis that eukaryotes evolved by the fusion of Archaeal and Bacterial genomes. Such trends of reductive evolution have also been uncovered by the analysis of protein fold and fold superfamily usage in the the three domains. An important finding from this study was that during evolution, Archaea were the first to adapt to stressful environments and in the processes were the first to start the reductive mode of evolution for adaptation. Hence Archaea was the first group to diverge away from a common ancestor leading to the tripartite world. It further suggested that the common ancestor was rich in terms of the protein folds it used and the genomic strategies. Its lifestyle resembled that of eukaryotes [124].

Interestingly such trends of reductive evolution were also observed in a survey of ribosomal protein (r-protein) gene families at the domain level [156]. The large number of protein familes common to Archaea and Eukarya, rather than Archaea and Bacteria meant an early divergence of archaea from the common ancestor. That implies that the precusrsors of modern ribosomes were complex and resembled more the eukaryotic versions. If that is seen it would be interesting to reconstruct the nature of ancestral ribosome to understand why it grew complex before getting into a reductive mode.

5.3 RNA World, Protein World or RNP World

Although the RNA World hypothesis was an attractive solution to the chicken-or-egg problem of whether RNA or proteins evolved first during the origins of life, it is largely unsupported. Without hard evolutionary evidence, RNA World is mostly presumptive and has yet significantly influenced the origins of life research [179]. Recently, hypotheses supporting the 'peptides first' scenario where a Polypeptide World evolved into the modern RNP world have been proposed [257]. These are advances of previously proposed hypotheses [204] and are finding support from phylogenomic studies [258].

Considering the "Ribosome" as the ultimate chronometer, if it is traced back to its simplest form it is more likely that there would be an RNP World. The "ribosome" is

emphasized, as it has a core that consists of both RNA and protein. rRNA is still a choice as a chronometer due its abundance and the economics of its use. However equally good phylogenies are obtained with r-proteins [259, 260]. Even before genetically encoded proteins existed, it is highly likely for an RNP world to have existed. It is easier to synthesize amino acids than ribonucleotides [261]. Although RNA folds spontaneously into its secondary structure, its tertiary fold is not very stable, especially when it is large. Both amino acids and RNAs bind to each other and almost always the 'induced fit' guides a higher level of assembly and stabilizes the structures [159].

An impressive range of ribozymes have been selected by directed *in vitro* evolution that are capable of catalyzing many of the biologically relevant processes such as nucleic acid replication, tRNA amino acylation, peptide bond synthesis [49]. Proponents of the hypothetical RNA World consider this as evidence for the RNA World. However, the rates of catalysis of these synthetic ribozymes are orders of magnitude lower compared to natural ribozymes or protein enzymes. There are no known natural ribozymes that catalyze similar reactions. Natural ribozymes are found associated with proteins [179]. Although the RNA components of such ribozymes have shown protein free activity *in vitro*, such activity is not observed *in vivo*. *In vitro* activity is observed under unusually high salt concentrations, which is not natural [155]. However, their catalytic rates increase many folds when associated with a protein. Remarkably similar results are obtained with *in vitro* evolved ribozymes. In the case of the L1 Ligase ribozyme discussed in the previous section, after selection for peptide dependent activity, a variant with $> 18 \times 10^3$ fold increase in activity was isolated. Both natural ribozymes and *in vitro* selected ribozymes that are RNP complexes supersede RNA-only ribozymes in catalytic efficiency. It is highly likely such primordial RNP complexes could have existed in a prebiotic RNP World but there is no evidence for an RNA World [179].

Since proteins predominantly perform most cellular functions and in vivo catalytic RNAs are always associated with proteins and dependent on them for their activity, the RNP system in all cells and especially in eukaryotes, are relics from a primitive RNP World [262]. It is also one of the features that support a complex common ancestor [263]., which has evolved by

simplification in Bacteria and Archaea while the Eukaryotic lineage retained it and further expanded it to a modern eukaryotic RNP World.

Furthermore, a recent study inferring the evolution of molecular functions from gene ontologies has uncovered metabolic origins of molecular functions [125]. Hydrolases, transferases, with (Nucleotide Triphospahe) NTPases and helicase activities were found to be the most ancient. Remarkably, helicases were among the first enzymes to use energy from NTP hydrolysis. In an earlier study it was shown that nucleotide metabolism is the oldest metabolic network [84]. Together these new reports suggest a "metabolism first" theory. Similar theories have also been proposed to explain the origin if the genetic code [264].

If a metabolism driven Protein World has taken over most of the functions why is translation still a RNP function? Perhaps the high degree of flexibility that is conferred by the RNA and the transient RNA-RNA or RNA-protein interactions that can be established is important [265]. Recognition of other molecules by base-paring interactions and the propensity and flexibility that RNA domains provide for large-scale movements have been selected over the course of evolution of the ribosome. Apart from a mechanistic perspective, RNAs could also have been selected as a means of cellular economy. Actively growing bacteria invest most of cellular resources in the protein biosynthetic pathway [183]. The cellular cost of synthesizing an RNA molecule is much lesser than that of synthesizing a polypeptide. Once synthesized a protein's maintenance and turnover is also relatively more demanding compared to RNAs. Thus maintaining a large portion of the protein synthetic machinery as RNA is perhaps due to selection pressures of the dynamics of the regulation of proteins biosynthesis.

5.4 Recruitment or Co-option is the most likely path to the origins of the ribosome.

The evolutionary path to the advent of translation is likely to be quite complex requiring multiple evolutionary novelties, important among them are a capacity to copy molecules and genetically encode products amenable to selection. Such innovations are envisioned as a natural outcome of

a primordial prebiotic polypeptide based chemistry that does not require an RNA World scenario for its continued evolution and selection [179, 204, 257]

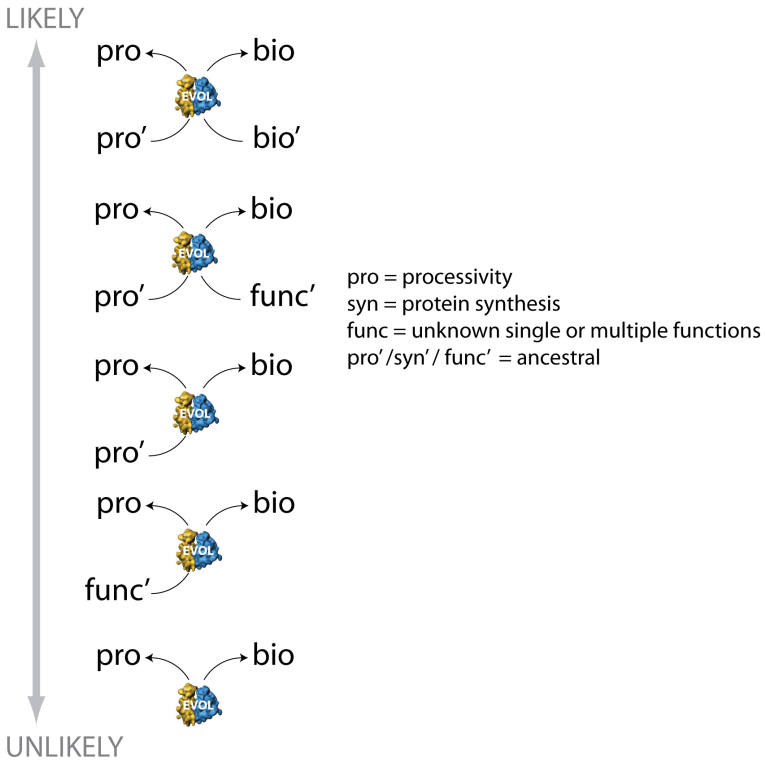


Figure 5.1: Possible scenarios of origins of ribosome and their likelihoods. Scenarios of de novo appearance of is complex is highly unlikely. The possibility of evolution of a complex such as the ribosome is most favored when a related primordial function or structure preexisted.

It is highly unlikely that a process as complex as translation could have arose in a single event of evolutionary novelty. It is equally highly unlikely that a multi-component molecular complex giving rise to the modern ribosome was a *de novo* evolutionary development. Since translation involves multiple stages and multiple players other than the ribosome, to satisfy the principle of continuity of evolution, many of the processes and the components must have preexisted. Instead the ribosome is likely to have gradually evolved from a much simpler and primitive complex, which was much smaller and perhaps had shorter RNAs and non-translationally synthesized polypeptides. Nonribosomal protein synthetases that do not use a template to synthesize proteins or their precursors could be very likely. Due to a high degree of similarity in functional strategies between the replication apparatus and the translation apparatus, recruitment of functions is most likely.

5.5 Main conclusions of the research

It has been possible for the *first time* to estimate the relative age of the ribosome comprehensively, including both SSU and LSU rRNA and r-proteins. The results show that

- The origins of the ribosome are not in peptide synthesis at the core of the LSU as generally argued due to the central role of the PTC in peptide synthesis.
- The origins of the ribosome lies in the core of the 'biological assembly' of the ribosome involved in the processivity of the ribosome, predominantly in helix h44 of the SSU which not only is involved in mRNA decoding and tRNA translocation but also crucial in forming most of the intersubunit bridges and mediating the signal relays between the two subunits.
- tRNA is at the heart of ribosome evolution. The functional domains of the ribosome have coevolved with the modern four-stemmed (clover-leaf) tRNA.
- r-proteins have coevolved with their respective rRNA domains from a very early stage of ribosome evolution and are crucial for the activity of ribosomes.
- Consequently, the division of labor between and RNA and protein components is indistinguishable. Therefore attributing relative importance of RNA over protein or *vice versa* with regard to a greater contribution to ribosomal functions is not reasonable.
- Both proteins and RNA are equally important to stabilize the structure and a RNA only proto-ribosome is unlikely to have existed.
- A primitive ribosome was most likely recruited to perform translation from related functions like replication. Alternatively elements of a replications apparatus could have been recruited to the translation due to similarity in the processes.

Chapter 6
Appendix

6.1 Evolution of rRNA in individual subunits of the ribosome

Trees representing the evolution of rRNA structure were reconstructed from either a balanced data set comprising equal number of sequences from species of the three superkingdoms of life or an unbalanced dataset. The unbalanced dataset is heavily biased towards bacterial sequences, which are all rRNA sequences available as DCSE alignments from the European rRNA Database (ErDB). The balanced dataset was used to avoid effects of sampling bias on tree reconstructions. The balanced dataset was limited to a total of 93 sequences with 31 sequences from each superkingdom of life. Only in the case of these species, both LSU and SSU rRNA sequences were available. This was important and required to reconstruct a universal tree of rRNA structural elements of both SSU and LSU structural elements. In case of the unbalanced dataset, a tree of substructures of either SSU only or LSU only had to be reconstructed when a corresponding SSU/LSU sequence was not available.

To compare rRNA helix ancestries between LSU-SSU and LSU only and SSU only trees, we calculated the relative age of each helix as a node distance (*nd*), the number of nodes from a hypothetical ancestor (root) in a relative 0-1 timescale (see Methods). These ages were traced in secondary and 3D structural representations of the molecules "evolutionary heat maps" and used to build timelines of development of components of the ribosome and their associated functions Only helices that are present in all three superkingdom, were included in the analysis.

Phylogenetic trees of SSU rRNA alone Figure 6.1, like the combined SSU and LSU rRNA

tree (Figure 2.5) shows SSU helix h44 is the oldest (*nd* = 0). This substructure, the penultimate helical stem in the SSU rRNA, is one of the most functionally important ribosomal substructures. It interacts with other SSU substructures responsible for mRNA decoding and with the LSU rRNA forming a functional relay [95]. Most of the interactions of the mRNA and the tRNA are hence centered in this helix. This relay is proposed to link processes in the SSU decoding site with LSU-based processes such as peptide bond formation and the release of elongation factors, thus modulating intersubunit interactions [95]. Helices h23, h24, h28, h30 and h34 are primordial (*nd* = 0.118-0.294); h23, h24, h28, h30 define the A, P and E sites of the SSU [95] and h34 is involved in tRNA translocation during the elongation cycle of translation [96]. However, some helices that are proximal to these ancient elements, such as helices h27, h29 and h31, are recent (*nd* = 0.471-0.912), suggesting they evolved after basic mechanisms were already established in the proto-ribosome, perhaps to refine established functions.

As with the SSU tree, the LSU tree Figure 6.2 is also congruent to the combined SSU and LSU rRNA tree (Fig. 2.5) and shows many functionally important regions are primordial. Helix H38 is one of the oldest substructures (*nd* = 0.0). It starts in the back of the particle, bends by about 90° and protrudes toward the SSU between domain V and 5S rRNA forming a crucial link between the two subunits [95]. Helices H73-H76, H89 and H90 that make up most of the catalytic core, the peptidyl transferase center (PTC) involved in peptide bond synthesis [99], are also ancient, (shaded yellow, *nd* = 0.267). The helical regions form the base of the polypeptide exit tunnel. In addition helices H2 and H7 of domain I (*nd* = 0.389), helices H26, H35, H35a, and H40 of domain II (*nd* = 0.6-0.9), helix H52 of domain III (*nd* = 0.648), and helices H61, H64, and H65 of domain IV (*nd* = 0.5-0.7) which are derived compared to helices of domain V, also comprise the peptide exit tunnel. Helices, H32 and H69 that directly interact with the SSU are also derived (*nd* = 0.74). As with SSU, not all substructures that are proximal to the functional center are primordial or follow a serial chronology. Derived structural elements were therefore added to a basic functional proto-ribosomal unit later in evolution. This suggests the proto-ribosome was able to perform its function, perhaps less efficiently, with a simpler structure.

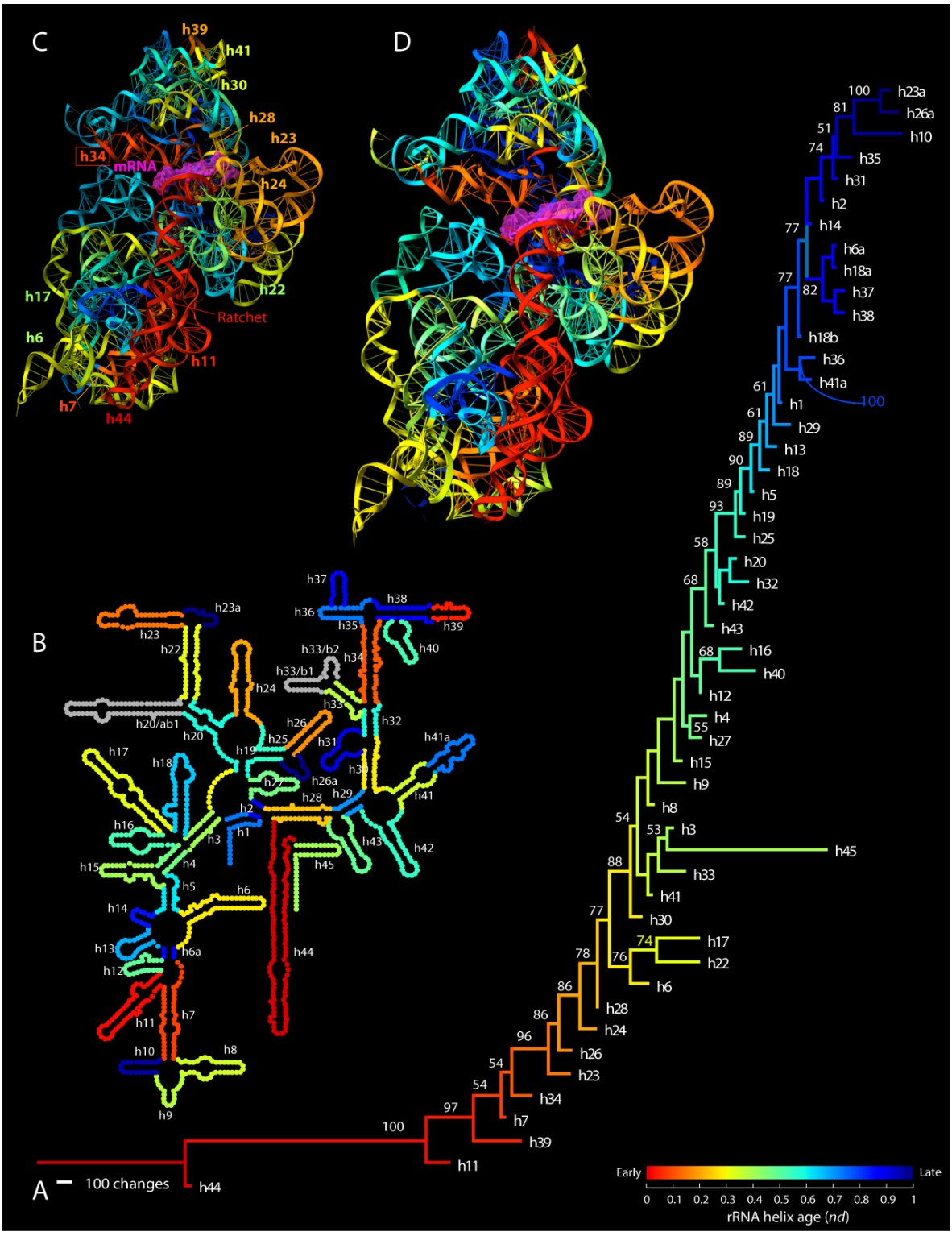


Figure 6.1: Relative age of SSU rRNA. A tree of SSU rRNA helical segments was reconstructed and relative age (nd) was determined. (A) SSU helices tree (18459 steps, CI = 0.348, RI = 0.809, HI = 0.652, g1 = -5.675) (B) SSU secondary structure '2D evolutionary heat map' based on *nd* from (A). (C) SSU '3D evolutionary heat map' derived from the combined LSU-SSU tree. (D) A similar 3D evolutionary heat map derived from a tree reconstructed of SSU helices alone. The 3D map shows the congruence of the SSU and SSU-LSU trees.

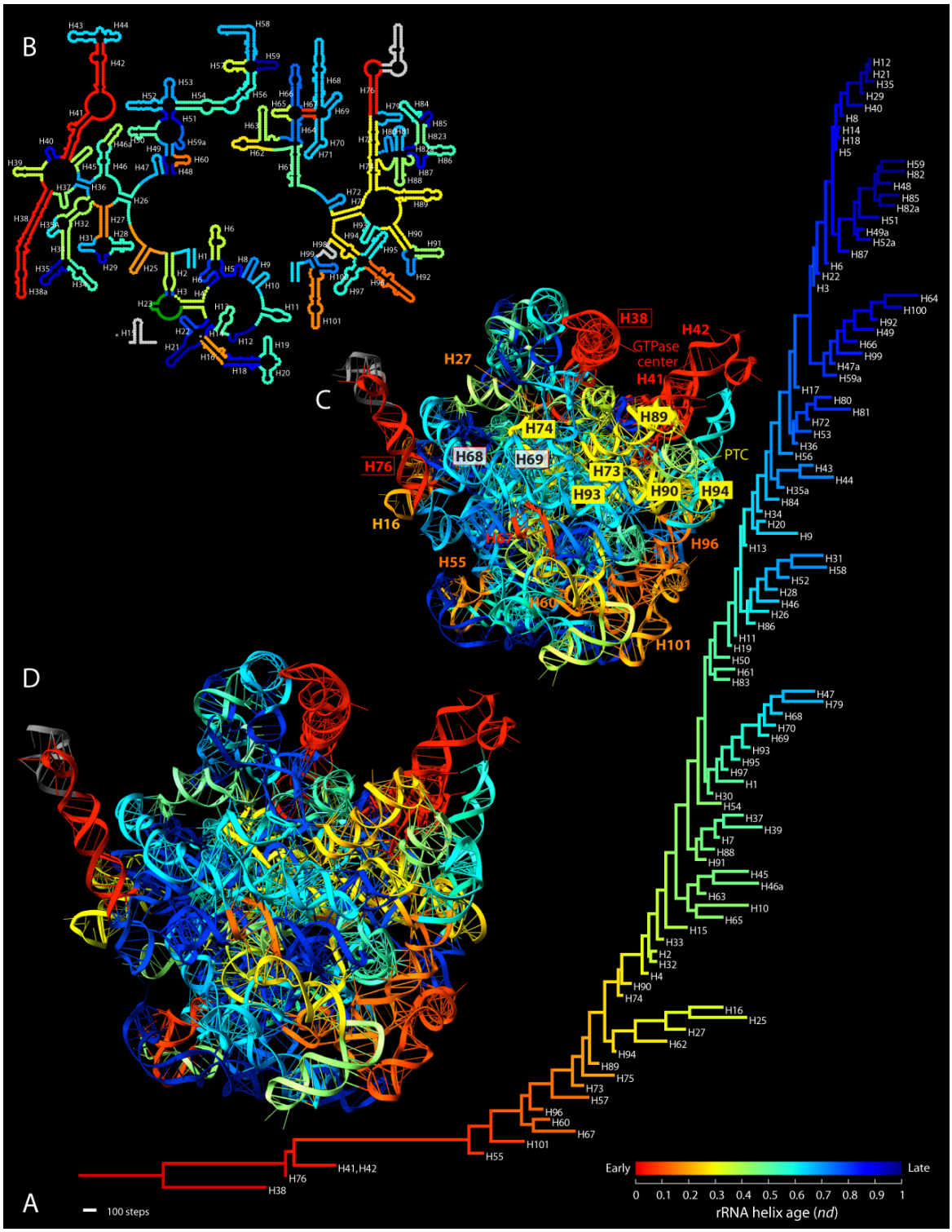


Figure 6.2: Relative age of LSU rRNA. A tree of LSU rRNA helical segments was reconstructed and relative age (nd) was determined. (A) LSU helices tree (11055 steps, CI = 0.226, RI = 0.714, HI = 0.161, g1 = -5.430) (B) LSU secondary structure '2D evolutionary heat map' based on nd from (A). (C) SSU '3D evolutionary heat map' derived from the combined LSU-SSU tree. (D) A similar 3D evolutionary heat map derived from a tree reconstructed of LSU helices alone. The 3D map shows the congruence of the SSU and SSU-LSU trees.

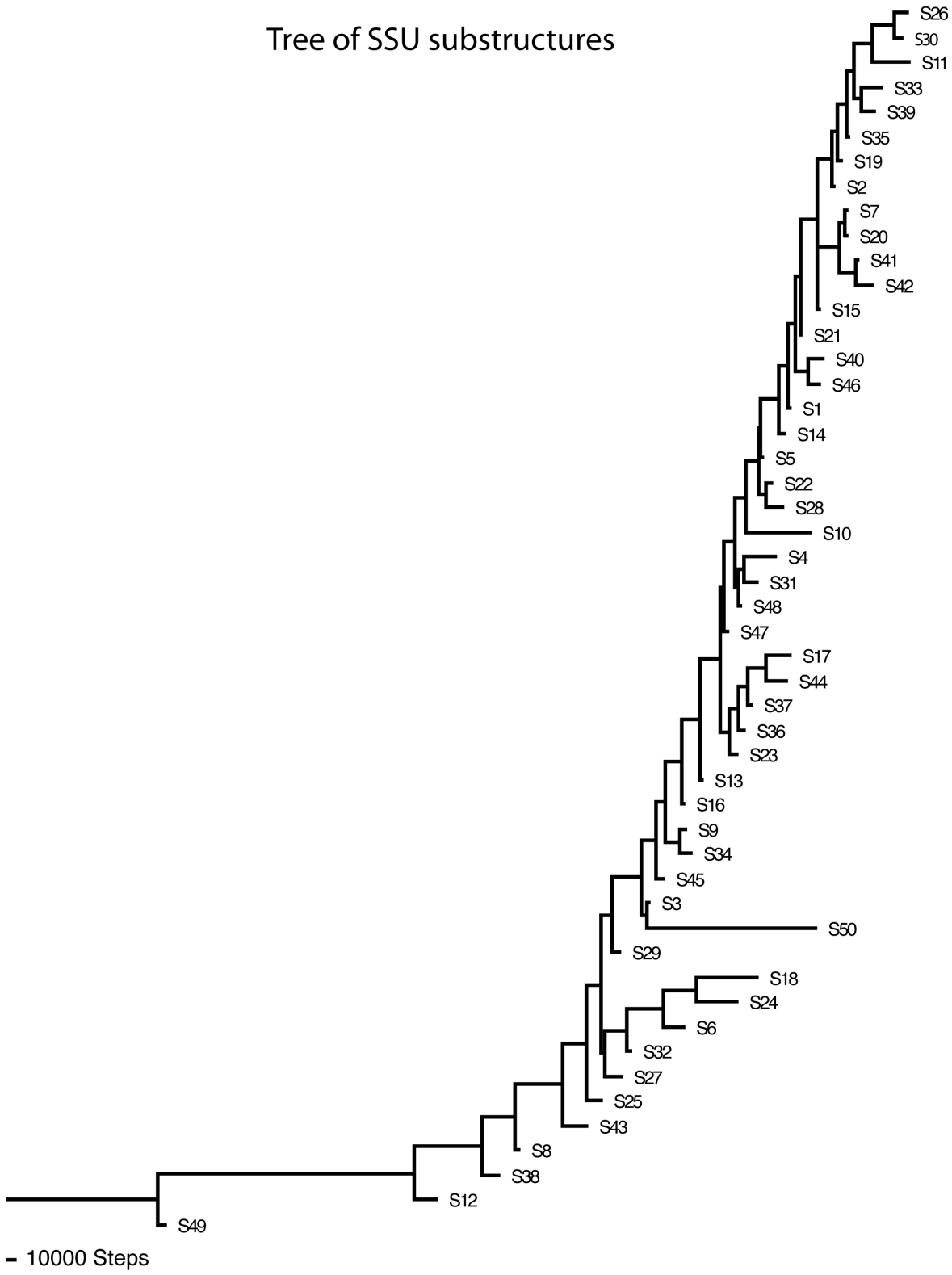


Figure 6.3: Tree of SSU helices reconstructed from ~19000 sequences from ErDB (Length = 39136steps; CI = 0.835; RI = 0.971; HI = 0.165; g1 = -192.782). The sampling is heavily biased towards Bacterial sequences.

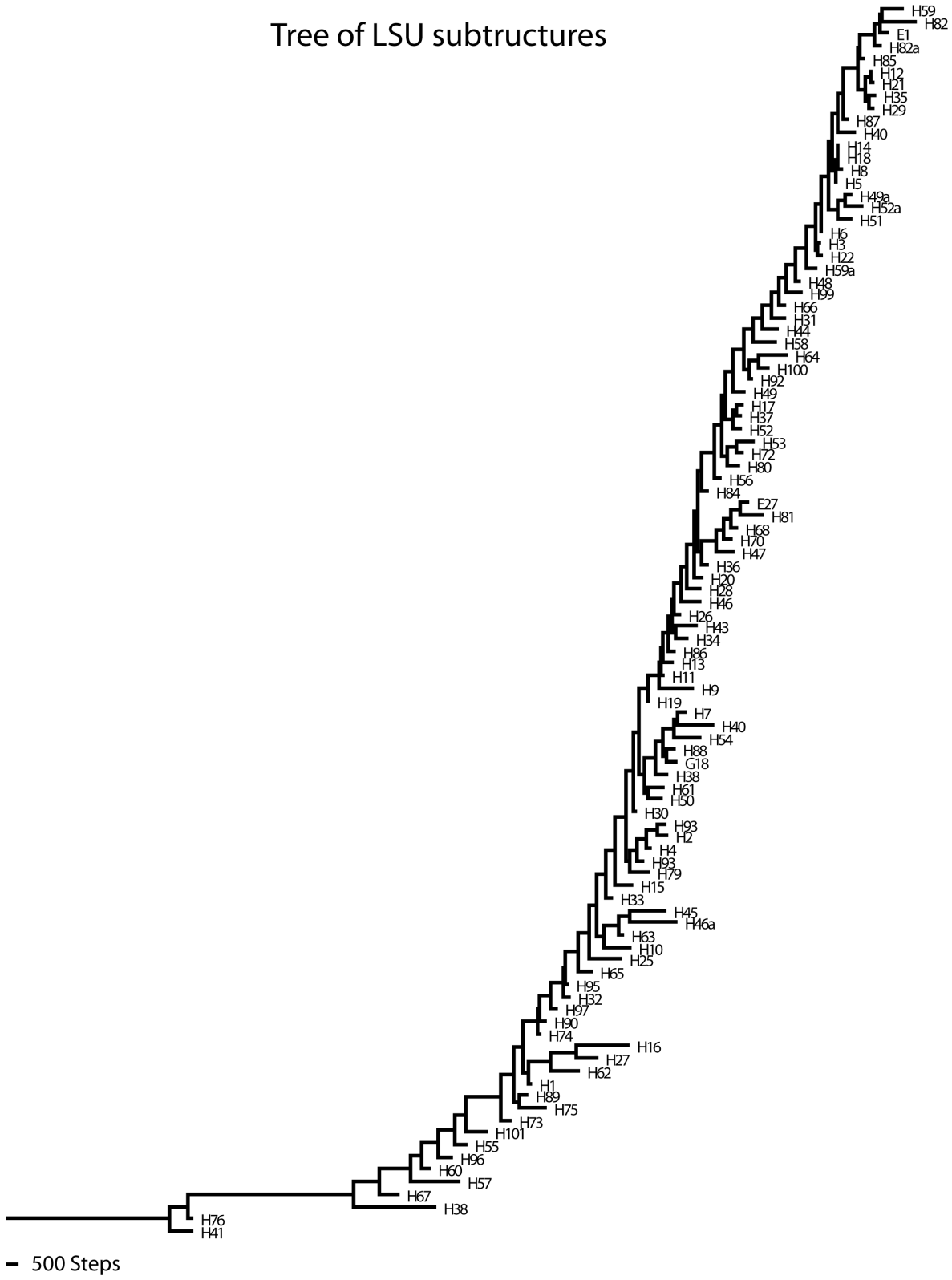


Figure 6.4: : Tree of SSU helices reconstructed from ~600 sequences from ErDB. The sampling is heavily biased towards Bacterial sequences. (Length = 138582 steps; CI = 0.265; RI = 0.751; HI = 0.735; G-fit = -24.507)

6.2 Conclusions

Comparisons of the combined LSU-SSU trees to either SSU or LSU trees show that the trees are congruent between the indvidual reconstructions. Combining LSU-SSU structural elements into one analysis does not change the relative palcement of the nodes in the reconstructed. Thus it shows the robustness of our method. The hypothesis put forth by both the individual and combined trees support each other and corroborate our arguments about the evolution of the early and primordial ribosome.

6.3 Materials and Methods

6.3.1 Data retrieval

The sequences of LSU and SSU rRNA were obtained from the European Ribosomal RNA database at (http://bioinformatics.psb.ugent.be/webtools/rRNA/) [62] in DCSE format in which the secondary structure of the rRNA is encoded in helix numbering lines for sets of alignments specific for Archaea (A), Bacteria (B) or Eukarya (E). Helix numbering lines identify the corresponding paired regions of each helix in the rRNA secondary structure. A total of 102 LSU and 102 SSU sequences were obtained. Since the database is biased towards bacterial sequences, a set of 102 sequences of both LSU and SSU from 34 species each from Archaea, Bacteria and Eukarya were used to reconstruct trees.

6.3.2 Determining relative age of rRNA structural elements

Since there are no explicit models for the evolution of RNA structure we limited our analysis to parsimony based methods implemented in PAUP* [131]. A novel phylogenetic method that reconstructs the history of molecular substructures of rRNA was developed earlier [11]. This method embeds structure directly into phylogenetic analysis [40]. Phylogenetic relationships were inferred on the basis of shared and derived characteristics in RNA structure using cladistic principles. Molecules were characterized by attributes that describe the topology of folded conformations. RNA secondary structures were first characterized using attributes that describe the overall ‘‘shape’’ (geometry) of molecules [133]. These attributes were then treated as linearly ordered multi-state characters that were polarized by fixing the direction of evolutionary

transformation toward molecular order. These trees describe a finite molecular system in which the 'leaves' represent the individual structural components of the molecule.

6.3.3 Character coding of RNA structure

RNA secondary structures are most suitable to study evolutionary relationships [134]. RNA secondary structures inferred from comparative sequence analysis were decomposed into structural elements (substructures) and their features (such as the length of stem tracts) were characterized using an alphanumerical format for cladistic analysis. Homologous components were treated as discrete entities and analyzed with maximum parsimony methods. Other alternatives are possible. In related studies, structural elements were characterized by their thermodynamic stability measured using their minimum Gibbs free energy increments [88]. These values were treated as discrete characters for maximum parsimony analysis. Coded characters were based on the length and number of double-helical stem tracts (S), hairpin loops (H), bulge and interior loops (B), and unpaired sequences (U).

In this study, topographic correspondence was the main criterion for determining character homology. It should be noted that unpaired nucleotides could form unusual base pairings or establish non-covalent interactions. These interactions are involved in high-order three-dimensional motifs like tetraloops, pseudoknots, A-minor motifs that stabilize RNA tertiary and quaternary structures are not considered in the structural models of this study. Several coding schemes are possible, however, character argumentation employed here is simplistic. That is, character coding disregards information and implications of higher order structure, coarse-graining its three-dimensional complexities into a simple framework of non-interacting helical segments and thus have avoided any bias to a given substructure. Our assumptions are corroborated by rRNA crystal structures. Nearly all of rRNA is helical or approximately helical, and RNA structure can effectively be considered a three-dimensional arrangement of helical elements [94]. Character coding relies however on correct prediction of secondary structure. Covariation based comparative sequence analysis has been successful in predicting structures with high accuracy of up to 96% [61]. Structural inaccuracies were therefore assumed not to be severe and were tolerated as systematic error, provided structures result from a same comparative sequence study or are folded using the same algorithm.

The coding of rRNA was based on secondary structure models for the large and small subunits inferred from sequences deposited in the ErRD and defined by comparative sequence analysis [59]. The SSU model contains 50 universal stem tracts (S) and several double-helical segments specific for Eukarya. The LSU model contains 100 universal stem tracts and several other stems specific to certain taxa. As described earlier the ribosome is essentially an arrangement of double helical stems. Thus helices (S) present in all three super kingdoms were used for the analysis. Note that universal stem tracts in these models are defined as those segments separated by multi branched or pseudoknot loops and are identified by numbers ordered in the 5'-to-3' direction. Character states were limited to 64, the maximum number accepted by PAUP* (http://paup.csit.fsu.edu/paupfaq/faq.html), and were represented by the numbers ***0–9***, case sensitive alphabets ***A-Z*** and ***a-z*** and special characters ***@*** and ***&.*** Structural features with longer than 64 nucleotide lengths were given the maximum state (&), and if missing, the minimum state (0). Structural alignments listed characters describing the structure in the 5'-to-3' direction as it is read in the sequence, and for each sequence segment, in the order S, B, H, and U. Stem tracts were defined by two complementary sequence segments and characters (named by a number and its prime) to account for the difference in nucleotide numbers between stem and unpaired segments. Helix numbering of the rRNA stems as in ErRD [59, 62] was used in the character coding and tree reconstruction exercises SSU helices are numbered S1-S50 and LSU helices are numbered A-I corresponding to the different domains. This was then reconciled with the standard Brimacombe numbering [63] used in the crystal structure of *Thermus thermophillus* ribosome [58].

The method was initially applied to 35 rRNA molecules sampled from all the three organismal superkingdoms of life and later extended to 102 molecules with equal representation from each superkingdom (Archaea, Bacteria and Eukarya). An in-house software module, MARTEN [135], was used to code characters from DCSE alignments and to generate executable files for PAUP*.

6.3.4 Phylogenetic analysis

The relative ancestry of rRNA structural elements were reconstructed using maximum parsimony methods in PAUP* v. 4.0-b10 [131]. The ANCSTATES command was invoked to define ancestral character states and polarity of character transformation. Phylogenetic trees were derived from heuristic searches using tree-bisection-reconnection (TBR) branch swapping and simple addition sequence. Phylogenetic reliability was tested by the nonparametric bootstrap method implemented using 5000 pseudoreplicates.

Since method used here produces intrinsically rooted trees, relative age (ancestry) of the individual elements could be determined by measuring the distance in nodes from the hypothetical ancestor (root) in a relative 0-1 time scale and is a measure of the total amount of independent evolutionary history, regardless of the tree topology [147]. Node distance (*nd*) counts the number of cladogenic events (nodes) along a lineage in the tree of rRNA helices starting with the first cladogenic event (bifurcation at the root) traversing to each terminal tip. Therefore the *nd* ancestry value of the oldest helix is 0 and 1 for the most derived helix.

6.3.5 Evolutionary Heat Maps

The 2D and 3D evolutionary heat maps were essentially generated as previously explained in chapter 2, under section 2.8.6

6.4 Data: Phylogenetic data matrices, sequences and alignments

6.4.1 PAUP data matrices

The input data matrix used to reconstruct either the SSU or the LSU tree of secondary structure elements are as follows.

```
Input data matrix for the SSU substructure tree

Dimensions ntax=74 nchar=102; Character state symbols = "0~9 A~Z a~z @&"


Note: The ErDB helix numbering is shown in the data matrix

                     1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node   123456789012345678901234567890123456789012345678901234567890123456789012345678
-------------------------------------------------------------------------------------------
s1           AAAAAAAAAAAAAA8AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA8AAAAAAAAAAAAAAAAAAAAAA
s2           8888888888888888888866888868888888888888888888888888888888888888A888889C9689CCCC
s3           KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKGKKKKKKKKKKKKKKKIKKKKKKKIKKKKKKKIKKFKKKKKKKGKEKIKIK
s4           IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGGGGGGGGGGGGGGGEEEGGGGGGGCGGGGG8888388888
s5           CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBCCCCCCCCC6CCCCC
s6           MMOOGKYOKMMMMMIIGIIIOOGGOIGIISSSKIcSMWWUOEUQYUMMKOaaYQWWYYWMIUUQYQYWEICEECCECC
s7           66666666666666666666666666666666666666666666666666666666666666666L666666666366666
s8           WWWWWWWWWWWWWWWWWWWWWWWWWWWWWSWWWWWWWWWWWWWUWWUWWWWUUUUUUUUUUUWWWIWWWWWKKKKAKIIIK
s9           KKKKIKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKIIKKKKIKIKKKKIKIIIIIKKKKIIIIKIKKKCCCCCCCCCC
s10          KKKKKKKKKKKKKKKIKIIIIKKKIIKKKKKKKKKKMMKKI6KKMKIM866666666666IMK0IIIIKGWGCACKACI
s11          KKKK4K4KGKKKKK444444664448444CC884O44GGK44C6CC444444IGGGGGGAG20C6A8AC8CG6GGGGI
s12          YYaaaaaaaYYaaaaaaaaaaaaaaaaaaaaaaaaaaaaWaaaaaaaaaaaaaaaaaWaaaa4WYaaaaaaW0WaWUa
s13          GGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGIGGGGGGGG0GGGGG
s14          AAAAAACAAAAAAACCCCCAACCCCACCAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAQAAAAACCCC0CCCCC
s15          868888888888668888888886888888888888888888888888888888888888888888B888888888088888
s16          IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIEIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIII9IIIIIGGGG0GGGGG
s17          ACECECGEECCCCEGGGGGGGGGCCGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGGGGGGGGGGEFGGGGGOQQMGOOOOQ
s18          KKIKIKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKeYOUUSGWYCYWQECCCCUSUSWWWUUGHUCUMY4424Y44442
s19          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC888888888888888888888888888888A88888CCCCCCCCCC
s20          46666666666666666666666666666666666666666666666666666666666666666F666666666E66666
s21          8888888888888888888888888888888888AAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAA8888988888
s22          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCFCACCCHBIBGHBBBB
s23          EEEEEEEEEEEEEEEEEEEEEEEECEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEHEEEEEKJKJKKJJJJ
s24          QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQWWWWWSUUSUUUSUUUUSSSUUSUQSQIUSSUUH2JCII2222
s25          SSUUQUSUUSSSSUUUUUUUUUUUUUUUUUUUUUUOOOOOKOOOOMOOOOMOOOOOOOOOOMOTOOOMOEPF8EGPPPP
s26          44444444444444444444444444444444444444444444444444444444444444444Z44444MFLFLNFFFF
s27          OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMOOOOOOOO4OOOOO4J4Q44JJJJ
s28          CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCPCCCCCNEODMNEEEE
s29          SOOSQQSSSSSSSSSSSSSSSSSSSSSSSSSSSSMMMMMMMMQMQQMMOOOOMMOOOMOMMOJKMMKOJcJDJJJHJI
s30          4444444444444444444444444444444444444444444444444444444444442444444I44444HZHJGHGEGF
s31          IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGEGGEEEEEEEEEEEEEEEEEEEEE4EGGGG4E4O44EEEE
s32          OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMOOOOOOODOOOOOECEAEECCCC
s33          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAEAAAAAEIEDDDHHHI
s34          KKIKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKEKKKKKECElDDCCCC
s35          8888888888888888888888888888888888888888888888888888888888888888G88888GEGGGGEEEE
s36          EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEECEEEEEBLBjAALLLL
s37          MMMMMKMMMMMMMMMMMMKMMMMMMMMMMMMMKGICGGGEGEEEEGAEGEEEEEEEEEEGEEHGIGGEELEHEELLLL
s38          UUSUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUUYYYYYYYYYYYUYYYYYYYYYWYYYYYYLYWYYYPIMBMYGGGI
s39          8888888888888888888888888888888888888888888888888888888888888888B88888BKBLBBKKKK
s40          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAABAAAAAS7SBSO7777
```

```
                    1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-----------------------------------------------------------------------------------------------------
s41         888888888888886888888888888888888666666666666666666666666666666L66666K7KMKK7777
s42         888888888888886666668866888668888866666666666666666666666666666J66666P8M6MY8889
s43         acccccccccccaaaaaaaYYaaaaYaaaaaaWSSSSSSQQQQQQSSQQSSKOQQQQQQMS6SOSSS6&6866LLLM
s44         GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG7GGGGG7&7655OOOO
s45         KKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKIKKKKKKKKKKKEIIIIKKIKKKKKKI7KKKKKCICEAAIIII
s46         AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACCAAAACAAAA8AAAAAAAAAAAAAAAAIAAAAAKHJOIJ8888
s47         EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEECEEEEEIGHHGHEEEG
s48         GGGGGGGGGGGGGGGGGGGGGGGGGGGEEEEGCEEEEEEEEEEEEEEEEEEEEEEEEECEEEEEBDBEAADDDD
s49         &@7@y@w&&@@&&&wuusws@@yyww@ss&&&&@&mk&&@yuywwwi&@&yy@@y@@@@swwJ&@&wyKqKKKKsrsr
s50         KK7KKKKKKKKKKKKKKKKKIKKKKKKKIIIIKKKKKKKKKKKKKKKKKIKAKKKKKKKKKwKKKKK@o@&&&qpqp
```

Input data matrix (continued):

```
                                  111
            78888888889999999999000
Taxon/Node  90123456789012345678901 2
-------------------------------------
s1          AAAAAAAAAAAAAAAAAAAAAAA
s2          888CCCC99CCCC8DDDDDCCCCC
s3          DDAKBKBKKKKKKKKKKKKKKKK
s4          6618AAA8888A888888888868
s5          666C9C9CCCCCCCCC6CCCCCCC
s6          88EGGEGCCCCGEKCCCCCCCCEE
s7          3336D6B66666666666666666
s8          AA8KRKUKIKKQKKKKKKKMMMKK
s9          000C0G0CCCCGAACCCCCCCCCC
s10         000K0G0ECEE0GSAAAAAAACEG
s11         000MP0UGGGE0EGEGEGGCCCGC
s12         000aUaUaaaaaaUaaaaaaaaaa
s13         000EAGEGGGGGGEGGGGGGGGG
s14         000CaCaCCCCCCCCCCCCCCC8
s15         0008G8G88888888888888888
s16         000ICGCGGGEIGGGGGGGGGGG
s17         000OASAQQQQMKMQQQQQQQQQ
s18         000EG4G44440244444444442
s19         330CCACCCCCCCCCCCCCCCCC
s20         0006E6F66666626666666666
s21         0008E8F88888888888888888
s22         FIC8686KKBBABIBBBBBBBBBB
s23         JJF9393HKJJGJKJJJIJJJJJJ
s24         IIJ6QZAGJ2222G2222222222
s25         66EJMiOHIPPNPCPPPOPPPONP
s26         HHGKQQQNOFFFFLFFFFFFFEDF
s27         444JGIG44JJIJ4JJJJJJJJJJ
s28         LOJEEEEQQEEDEOEEEEEEEEEE
s29         IJHHHHHJJJJGLJJJJJJJJJKLI
s30         GHF@&MqHHGGCIGGGGGGGGHIF
s31         444BBBB44EEDE4FFFFFEEEEE
s32         EECGFFFEECCCCECCCCCCCCCC
s33         EEEA8AADCHHIHDIIIHIGGHIH
s34         DDDTTRTDCCCCCECCCCCCCCCC
s35         FFEFFFFGGEDEEGEEEEEEEEEE
s36         AAA8878AALLKLBLLLLLLLLLL
s37         EEEABBBEELLJLELLLLLLLLLL
s38         AAALKKKMOIIGIVIIIIIIIIII
s39         BB9HHHHBBKKJKBKKKKKKKKKK
```

```
                                  111
            78888888888999999999000
Taxon/Node  90123456789012345678901 2
-----------------------------------
s40         SSSKKIKSS7767M7777777777
s41         HHI8888KK7767K7777777777
s42         AAAA999MO8888V8888899988
s43         666776766LO4t6NNNNMNNLPO
s44         667Ko@o77MP5u7OOOONPPNQP
s45         DDCCEFECCGHAICIIIIIIIIII
s46         KHLm8CKLK6G08J8888888888
s47         JGJFHIHJIGJ9FHFFFEFDDEGF
s48         BBBB9BBAADADDADDDDDDDDDD
s49         KKHKJJJKKqKbpKprrrroopop
s50         uq&&&&&&&o&Zn&nppppmmnmn
```

Input data matrix for the LSU substructure tree

Dimensions ntax=120 nchar=102; Character state symbols = "0~9 A~Z a~z @&"

Note: The ErDB helix numbering is shown in the data matrix

```
                    1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-------------------------------------------------------------------------------------------------------
A1          GGT44GEEEEEEEEPMMOAEEGGGEESSQSAEGEGGGNGGGGGGGGGGGGGGUGGKNGGFGGGNGGGGG0JHHJJKJJH
B1          IKKKKIIIKKKKKIAAAAKKKKKIIIIIIIIIIKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKIKKKKKKK
B10         EEEEEEEEEEEEEE0000EEEEEEEEEEEEEEEEEEEEEEEEGCDEEEEEECEEEEEECCCEEEEEECEEEEEEEE
B11         444444444444400004444444444444444444444444444444444444444444444444464444444444444
B12         EEEEEEEEEEEEEE0000EEEEEEEEEEEEECEEEEEEECECCCCEEEEEEEEEEEE6EEEEEEEECCCCCCCCCC
B13         666666666666600006666666666666666666666666668666666666666666666666666666666666
B14         MGIOOEKKKKKKK0000CEEMMMKKKKKKKKOGKGKIGGIKIIGICIIIIIIEGGGEEEEEGIIIKKKOGIIGIGGM
B15         OSSWWOaaaaaQQa0000QQKQQQaaaaaaaSQWOOMUKKQKYUQYI6666QQQMMMSSWQOQOOOOQQGMEY0QaaS
B16         AAACCAAAAAAAAA0000AAAAAAAAAAAAAAAAAAAA6A8AAAA6AA8888AAA8A6AAA8AAAAAAAAAAAAAA
B17         666666666666660000666666666666666666666666666666666666666666466666666666666666466
B18         EEEEEEEEEEEEEE0000CCCEEEEEEEEEEEEEGGGGEEEGGEGGGGGGGGGGEEGGGGGGGGGGGEEEEEEEEEE
B19         AAAAAAAAAAAAAA0000AAAAAAAAAAAAAAAACCCACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
B2          888888888888884444888888888888888888888888888888888888888888888888888888888888
B20         444444444444400004244444444444444444444444444444444444444444444444444444444444
B21         688888688888880000666888888888888888888888888688888468888688888866688888888888
B3          IKKKKKKKKKKKKK9999KKKKKKKKKKKKKKKKMMMMMMMMMMKMMMMMKMMMMMKMMMMMKMMKMIIIIIIIIII
B4          666666666666666666666666666666666666666666666666666666666666666666666666666666466
B5          666666666666666666666666666666666666666666668688888888888888888888888888888886
B6          IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIKKKIIIKKKKKKKKKKKKKKKIKKIKKKKKKKKKAA8688A8A4
B7          466666666666666666666666666666466666666666666666666666666666666666064444446464
B8          AAASS48A888AAI1111000AAAIAAA6AA6AGEAEEECC0EECEEC&CCC8EEGGGAEEC0EEEGGCICIIIGGGO
B9          AAAEC8GKKKKCCI0000AA6GGGACKKMKM8AIE868IIIGIIIIIIKIIIIIII68GIIIIG6KCCUOUaUcUSUM
C1          NOOQQOommmmOOo2222PQPOOOqmkmokkWmoMMIIGGGIIIGIKEMQQIIIOMQOMKKIIMSGIKMMMOESBGBQ
D1          EEEEEEEEEEEEEE7777EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE7EEEEEGCBACe5C4C
D10         4444444444444444444444444444444444444444444444444444444444444444444444444444L4d4
D11         CCCCCCCCCCCCCC9888CCCCCCCCCCCCCCCCCCCCCCCCCCCCAACCCCCCCCCCCCCCC6CCCCC9EG6E6KDOD
D12         AAAAAAAAAAAAAAEDDDAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAARAHA
D13         IIIIIIIIIIIIIIYXXXIIIIIIIIIIIIIIIIIEIGGGGGGGIGIIIIGGIIIIIGGGIIIIIGGGG687889E7A7
D14         wuswwwyyyyyyuuwXWWWw@ywwwwyyyyyyyuwsccccceecaeYccWccccccccacWYecccYaUUacokommoKmFo
D15         KKKKKKKKKKKKKK4444KKKGGGKKKKKKKKGKKKKKKKKKKKKKKKKKKGGKKKKKKKKKKKKKKK000000C090
```

```
                    1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
----------------------------------------------------------------------------------------------------
D16         66666666666666CGGG66666666666666666666666666666666666666666666666666666666000000A0B0
D17         &&&&&&&&&&&&&&GGGG&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&A09000V0E0
D18         CCCCCCCCCCCCCCKKKKAACCCCCCCCCCCCCCEEEEEEEEEEEEEEEEEEEEEECAEEEEEEEEE000000A0V0
D19         88888888888888IIII88888868888888888888888888688888888888888888888888F00000W0Y0
D2          SSSSSSSSSSSSSSIGIHOSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSBSES
D20         CCCGGCIKMMMCCMFFFFEEEKKKKMMMMKMCKKOAI8GGIAEGGECGIGGGCEEMEE0IGG0GIKGODAH090G9G9
D21         EEECCECEEEEEEEFFFFEEAEEEEEEEEEEEEEECEEEEEEEEEACCCCCCCEEEECCCC0CCCAE&F7AFCFF7F
D22         MMMMMMMMMMMMMEEEDMMMMMMMMMMMMMMMKKMKMMMMKKKKMKMMMMMMKMMKKMMM0MMKKM8bAIcDccAc
D3          EGGIIEEEEEEGGEAAAAEEGEEEGEGGEGEEEGCEGGGGGGCGECGCCCCAAACCAAAAAAGGGGCCIGCEGI0I0E
D4          44444444444444GGGI4444444444444444444444444444224442444444444444444444424404E4
D5          EGGKKCKKMMMCCIAAAAEKEMMMKKKKKKKCGM88888888888888888888888888886888M8GSQUEQCO
D6          GGGGGGGGGGGGGBBBBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGKGLG
D7          KKKKKKKKKKKKKKGGGGKKKKKKKKKKKKKKKKKKKKKKKKKKOKKKKKKKKKKKKKKKKIKKKKKKKKKKMKaK
D8          KKKKKKKKKKKKKKCCCCKKKKKKKKKKKKKKKKIIIIIEIIIIKIIIIIIIGIIIIIGIIIIIIIIIIKKKKKEK4K
D9          CCCCCCCCCCCCCC8888AAACCCCCCCCCCCCCEECECCCEEEEEEEEEEEEEEECECEEEEEEEEECACCCC4CEC
E1          44444444444444JJJJ4444444444444444444444444444444444444444444444444444F9dJ9B99fA
E10         AAAAAAAAAAAAAADDDCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAFAAAAA77G77977G7
E11         GGGGGGGGGGG44GFFFFGGGGGGGGGGGGEGGIIMM888QMIIIGECGGKKMMMKKMMMMBMMMMM8E6PG9GGJF
E12         WkkggkkkkkkWWkAAAAUUUkkkkkkkkkkkkikWSSSOOOWWWUWGUQUUSUWWWWWWWWWAUUOWWMA5N9C99J9
E13         CCCCCAC0KKK88K99998CCAAACCCCCCC8ACCACCCCAAAAAGAAAAAAAAA8CAAAAACACE6E92CG6864M6
E14         MQQQQQQQQQQQQQCCCCOOQQQQQQQQQQQQQQeaYYSSQSaaWaU0000WYUYUYUGccW9aiUWYD8A8Q4MG9A
E15         2EEEEEEEEEE22E8888222EEEEEEEEEECEE08CC668A8A8E60000AA688EGEAAABEE88C6AD49899D9
E16         44444444444444BAAA4444444444444444444444444444444444444444444444444444IKHDK6MKHL
E17         86688688888888DDDD66688888888888888688666888868888888888668888E66688OIEIMCMMEK
E18         YYYYYYYYYYYWWYCCBCYYYYYYWYYYYYYYYYYWYYYYYYYYYYYYYYYYYYYYKYYYHYYYYY8FEAERDEDN
E19         GGGGGGGGGGGGGGLLKLGGGGGGGGGGGGGGGGGGEGGGGGGGEGGGGGGGGGGGGGGOGEGGG4N6MOGOO6O
E2          CCCCCCCCCCCCCCwyyyCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC0CCCCCEcIed9ddMd
E20         IUSIISSSWWWKKWLLKLIIIKKKWSSSSSSIKWWC8GWWSEMCEIUOOOOSQMUOQC0QSUKCOIOWC76B7H7757
E21         IIIIIIIIIIIIIICCBCIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGAIIIDIGGIIDAMCAKAALK
E22         8AAAAAAAAAAAAA6666888AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA4AAAAAsKG6K8JKGK
E23         KKKKKKKKKKKKKK4444KKKKKKKKKKKKKKKKKKKKKKKKKKKMKIKKKKKKKKKKKKIKKK9KKKKKsGP8GKGGkG
E24         88866888888888DDDD88888888888888886888888888888888888888888888988888FZIEdCdbFQ
E25         WWWWWWWWWWWWUWW6666WWWWWWWWWWWWWWWYcceccagccccaYYYYYYYceceeeccLaceYkASWOXUWVAO
E26         CCCCCCCCCCCCCCLLLLCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC0CCCGCCCCCKAATAVAAKV
E27         CCCCCCCCCCCCCCGGGGCCCCCCCCCCCCCACCCCCCCCCCCCCCCCCCCCCCCCCCACCCOCCCCCKKKUKAKKKA
E28         CCCAACCCCCCCCCGIIICCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC0CCCECCCCCMKKAJKKJMK
E3          88888888888888EEEE88888888888888888888888888888888888888888888H88888CJMBHFIHMH
E4          88888888888888DCCC88888888888888888AAAAAAAAAAAAAAAAAAAAAAAAAAA5AAAAAGEJ7DmHDJF
E5          66666666666666DDDD66666666666666666666666666666666666666666666366666J9E99H98F9
E6          GEEGGGGGGGGGGG6666GGGGGGGGEEGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGG0GGGGG9EGhFFHFHF
E7          666666666666666jjjj66666666666666666666666666666666666666666666666666BCNDCDCANA
E8          GEEGGGGGGGGGGGDDDDGGGGGGGGGGGGGGGAAAAAAAAAAA8AAAAAAAAAAAAA8AAA6AAAAAD95DADBB5C
E9          46666666666666AAAA44466666666664646666666666666666666666666646660666668 9CG9G99C9
F1          AAAAAAAA888AAAEEEEAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA8AAAYAAAAAABBMABBAAB
G1          QQQQQQQOQQQQQQSRRSOOOQQQQOOOOOOOOQQQQQPQQQOQQQQQQQQQQQQQQQQJQQQHQQQQQEFFHEGFEEG
G10         GGGGGGGGGGGGGGaaaaGGGGGGGGGGGGGGGGG5GGGGGGGGGEGEEGGGGEGG7GGGEGGGGGCDDHD9DCCJ
G11         CCCCCCCCCCCCCC6666CCCCCCCCCCCCCCCCCC2CCCCCCECCCCCCCCCCCCC0CCCHCAACEC88L8L88CL
G12         44444444444444HHHH4444444444444444444H444444444444444444440444B444443CC8C8CC38
G13         EEEECEEEEEEEEEAAAAEEEEEEEEEEEEEEEEECEEEEEEEEEEECEEEEEEE0EEE3EEEEEEAA7A7AAC7
G14         66666666666666333366666666666666666663666666666666662666660666D666664EEAEAEE4A
G15         KKKKKKKKKKKKKCCCCKKKKKKKKKKKKKKKIIICIIIIIIIIGIIIIIIIIIII3III4IIIIIEAA8A9AAE9
G16         OOOOOOOOOOOOOFEEEOOOOOOOOOOOOOOOOOOFOOOOOOOOOOOOOOOOOOO4OOOAOOOOOB66E6E66AE
G17         MMMMMMMMMMMMMHHHHMMMMMMMIMMMMMMMMMMMGMMMMMMMMMBMMMMMMMMMM0MMMEMMMMMKDD4D4DDJ4
G18         IIIIIIIGIIIIII4664IIIIIIIIIIIIIIIGIII6IIIIIIIIIIIIIIGIIIII0IIIBIIIIICGI5I5GIC5
G19         88888888888888CCCC8886668888888868AA4BAAAAAA8AAAAAAA8AAAAA3AAAAAAAAAMCCICICCMI
G2          MMMMMMMMMMMMMIHHIMMMMMMKMMMMMMMMKMMMDMMMMMMOMMMMMMMKMMMMMAMMMDKMMMMJBBDBEBBJE
G20         EEEEEEEEEEEEEECCCCEEEEEEEEEEEEEEEEEECEEEEEEEEEEEEEEEEEE7EEEOEEEEEIMMMMLMMIM
```

```
                    1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-----------------------------------------------------------------------------------------------
G3          000000000000008888000000000000000000C00000000M0000000000000000B000008eXHYIhY8I
G4          y&&&&&&&sss&&&CCCC@@@&&&y&@&&y&@&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&C&&&&&AIICIDIIAD
G5          666CC66688888CNNNN666888A8AAAAC68CGGA2GGEGGEGGGGGGGGGGGGGGGGGBAGG8IJaT9UQdULR
G6          AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA0AAAAAAAAAAAAAAAAAAAAAAAAOAAAAACJEND&FGEm
G7          AAAAAAAAAAAAAANNNNAAAAAAAAAAAAAAAACCC0CCCCCCCCCCCCCCCCACCCCCCChCCCCCDEAuEEDGHI
G8          22222222222222EEEE22222222222222222222222222222222222222221222c22222FBCAH7EGFB
G9          44444444444444dccc444444444444444444844444444444444444443444E44444CBBIBABBCK
H1          MOOOOOOOOOOOOLLLLMOOOOOOMOOOOOOOOOOOMOOOOOOOOOOOOOOOOOOOO3OOOBOOOOOMBAIAKAALK
H2          GGGGGGEEGGGGGGDDDDGGGGGGGEEEEEEGEGEEEEEEEEEEEEEEEEEEEEEE0EEEGEEEEEJIIAIAIIJA
H3          UUUUUUUUUUUUUUFFFFUUUUUUUUUUUUUUUUWWWGWWWWWWWWUWWWWWWWWWW5WWWEWWWWWHJJIJIJJGI
H4          GGGCCGGGGGGGGGEEEEGGGGGGGGGGGGGGGGGGEGGGEGGGGCGGGGGGGEGEE3GGGPGGGGGLHGJGJFGJJ
J1          88O888868888888QPPP88888888OOOO8888888N8868886888888888I886488J88888A8EDJDGGGD
J2          8A7AA888AAAAAAGFFGAA8AAAAA77778A88AAAGAAAAAAAAAAAAAAAAAA78A0AAA8AAAAAOJOMT8KQKQ
J3          agAggcggkkkcccAAAAYYYccceeAA8AgcccUUUAUUUUUUUUUUUUUUUUUASU0UUUAUUUUU6679777757


Input data matrix for the combined LSU-SSU substructure tree


Dimensions ntax=151 nchar=93; Character state symbols = "0~9 A~Z a~z @&"


Note: The ErDB helix numbering is shown in the data matrix


1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-----------------------------------------------------------------------------------------------
S1          AAAAAAAAAAA8AAAAAAAAAAAAAAAAAAAAAAAAAAAA8AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
S10         KKACK66KAKK6I6AKEKKMAM60GKGCKK6GKMMA0AI6IIKII0KEKKKKKCIK60AACSG8WKKK66EEIIMAKK
S11         KKEGK4IOG4CGC462GGG4E6GPCKEGMMG0444C0GIG444440GGC8644GA4G0CGCGC48GCC44GE6CGGKK
S12         YYaWaaaaaaaWa0aaaaaaaaUaaaWaaaaaaaa0aaWaaaaa4aaaaaaaaaaaaaaaUaaaaaaaaaaYaaWYY
S13         GGGGGGGGGGGGGG0GGGGGGGGAGGGGEEGGGGGG0GGGGGGGGIGGGGGGGGGGGGGGGEGGGGGGGGGGGGGGG
S14         AACCAAAACCAAAA0ACAAACAAaCACCCCACAAAC0CCACCCCCQCCACAACCACACCCCC8ACAAAAACCAAACAA
S15         868888888888880888888G88888888888080888888B88886888888888888888888888888888
S16         IIGGIIIIGIIIII0IGIIIGIICGIGGIIIGIIIG0GGIIIIII9GGIIIIIGIIIIGGGGGGGIIIIIGEIIIGII
S17         ACQMEGGGQGGGGGGEQGGGQGGAOCKOOOGSGGGQ0QQGGGGGGFOQGGGGCQGGGMQQQMQGQEGGGGQQGGGOCC
S18         KK44ICUe4KYSUCYG4UUQ4CSG4K24EEW4WYO4042WKKKKKH44YKKKK4UKU04444ZE4KKKCC44CWU4KK
S19         CCCCC888CC8888C8C888C88CCCCCCC8A888C3CC8CCCCCACC8CCCCC8C8CCCCCC8CCCC88CC888CCC
S2          88D68888D88888889888D88C98C9CC8C888C8DC888888ACC886689888CCDC8C8C88888CC888C88
S20         46666666666666E66666666E6666666666666066666666F6666666666666662666666666666666
S21         88888AAA88AAAA9A8AAA8AAE888888A8AAA8088A88888C88A88888A8A888888A8888AA88AAA888
S22         CCBBCCCCBCCCCCGCKCCCBCC6HCBH88C8CCCBFBBCCCCCCFBBCCCCCKCCCABBBIBCBCCCCCBBACCBCC
S23         EEJJEEEEJEEEEEKEHEEEJEE3KEJK99E9EEEJJIJEEEEEEHJJEEEEEKEEEGJJJKJEJEEEEEJJEEEJEE
S24         QQ2CQUSQ2QUSUUIQGWWU2SUQHQ2I66UZUWW2I22SQQQQQI22UQQQQJQQS2222G2S2QQQUU22SUS2QQ
S25         SSP8UOOOPSOOOMEOHOOOPOOMEUPGJJOiOOOP6OPOUUUUUTPNMUUUUIOUONPPOCPOPUUUOOPPOOMPSS
S26         44FF4444F44444L4N444F44QM4FNKK4Q444FHFF444444ZFD44444O444FFFELF4F44444FF444F44
S27         OOJQOOOOJOOMOO4O4OOOJOOG4OJ4JJOIOOOJ4JJOOOOOO4JJOOOOO4OOOIJJJ4JOJOOOOOJJOOOJOO
S28         CCEDCCCCECCCCCMCQCCCECCENCENEECECCCELEECCCCCCPEECCCCCQCCCDEEEOECECCCCCEECCCECC
S29         SOJDOOMMJSQMKOJOJMMMJMOHJSLJHHOHMMMJIJIMSSSSSJJLQSSSSJMSOGJJKJIMcSSSOOJJMQMHSS
S3          KKKKKKKKKKKKKKKEKKKKKKIKBKKKKKKKKKKKKDKKKKKKKKFIKKKKKKKIGKKKKKKKKKKKKKIKKKKKKKK
S30         44GJ4444G44444G4H444G42&H4IH@@4M444GGGF444444IGI44444H444CGGHGF4Z44444GG444E44
S31         IIFOIEEGFIEEEE4E4GGEFEEB4IE4BBEBGGGE4FEEIIIII4EEEIIII4EIEDEFE4EEEIIIEEEEGEEEII
S32         OOCAOOOOCOOOOOEOEOOOCOOFEOCEGGOFOOOCECCOOOOOOOCCOOOOOEOOMCCCCECOCOOOOOCCOOOCOO
S33         AAIDAAAAIAAAAADADAAAIAA8EAHDAAAAAAAGEHIAAAAAAEHIAAAAACAAAIGIHDHAIAAAAAHHAAAHAA
S34         KKClIKKKCKKKKKDKDKKKCKKTEKCDTTKRKKKCDCCKKKKKKECCKKKKKCKKKCCCCECKCKKKKKCCKKKCKK
S35         88EG8888E88888G8G888E88FG8EGFF8F888EFEE888888GEE88888G888EEEEGE8E88888ED888E88
```

```
                  1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-----------------------------------------------------------------------------------------------------------------
S36         EELjEEEELEEEEEAEAEEELEE8BELA88E7EEELALLEEEEEECLLEEEEEAEEEKLLLBLELEEEEELLEEELEE
S37         MMLHMEEGLMEEGGEEEGGGLEEBEMLEAAEBGICLELLEMMMMMHLLEMMMMEGMEJLLLELALMMMEELLIEELMM
S38         UUIBSYYYIUYYYYMYMYYYIYWKPUIYLLYKYYYIAIIYUUUUULGIYUUUUOYUYGIIIVIYIUUUYYIIWUYGUU
S39         88KL8888K88888B8B888K88HB8KBHH8H888KBKK888888BKK88888B888JKKKBK8K88888KK888K88
S4          II88IGEG8IGEGG3G8GGG8GGA8I8888GAGGG8688GIIIIIC86GIIII8GIEA88888G8IIIGG88GGG8II
S40         AA7BAAAA7AAAAASASAAA7AAKSA7OKKAIAAA7S77AAAAAAB77AAAAASAAA6777M7A7AAAAA77AAA7AA
S41         887M8666786666K6K6667668K87K88686667H77686888L7768888K6866777K7678886677666788
S42         88868666886666M6M6668669P88YAA696669A89666666J8868886O6668989V8688886688666888
S43         acN8cQKSNcQOSQ6S6SSSNQQ76ct677Q6QSSN6NMQaaaaa6LPQaYYa6QaQ4NML6OS&caaSSLOOQMLcc
S44         GGO6GGGGOGGGGG5G7GGGOGGo7Gu5KKG@GGGP6OOGGGGGG7OQGGGGG7GGG5PNN7PG&GGGGGMPGGGOGG
S45         KKIEKIKKIKKKKIAICKKKIKKECKIACCKFKIKIDIIKKKKKK7IIKKKKKCKKIAIIICIEIKKKIIGHKKKIKK
S46         AA8OAAAA8AAAAAIALAA88AA8KA8JmmACCCC8K88AAAAAAI88AAAAAKAAA0888J8AHAAAAA6GAAA8AA
S47         EEFHEEEEFEEEEEGEJEEEFEEHIEFHFFEIEEEDJEGEEEEEECEGEEEEEIEEE9DFEHFEGEEEEEGJEEEEEE
S48         GGDEGEECDGEEEEAEAEEEDEE9BGDABBEBEEEDBDDEGGGGGCDDEGGGGAEGEDDDDADEDGEEEEDAEEEDGG
S49         &@pK7@@&rwy@&&KwK&&irw@JK@pKKK@JumkoKrr@uwuwsJsoww@@yKssyborpKp&q&&&yyqK@wwr@@
S5          CCCCCCCCCCCCCC6CCCCC6CC9CCCCCCCCCCCC6CCCCCCCCBCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
S50         KKn&7KAKpKKKKK&K&KKKpKK&@Kn&&&K&KKKmuppKKKKKKwqmKKIKK&KKKZmpn&nKoKIIIKo&KKKpKK
S6          MMCEOKYcCYUQQOEUCWWMCQWGEOECGGYEESMC8CCYGIIIIUCEYIOGGCMIWGCCCKEMIKSSaaCCYUIEMM
S7          66666666666666366666666D66666666666636666666L66666666666666666666666666666666
S8          WWKKWUUWKWWUWUAWKWWWKUURKWKKKKUKWWWMAKKUWWWWWIIKWWWWWIWSUQMKMKKWKWWWUUKKWWWIWW
S9          KKCCKIIKCKKIKKCICKKKCIK0CKACCCKGIIIC0CCKKKKKKICCKKKKKCIKIGCCCACKCKKKIICCIKICKK
A1          EEKHEGUGKEGGGGJGKGGGKGFE0EHJ99GBGGNJEKHGMPOEAGKHG4TGGIKEG0JKGHHNJEGGGGKKGGNJES
B1          IIKKKKKKKKKKKKKKKKKKKKKKKIKKIIKKKKKKIKKKAAAKKKKKKKIIKKKKKKKKKKKKIKKKKKKKKKII
B10         EEEEEEEEEEEEEEEEEEEEECEEEEEEECDEEEEEEC000EEEEEGEEEEEEECEEEEEECCEEEEEEEECEEEE
B11         44444444444444444444444444444424444444400044444444444444444444444444644444
B12         EEACEEEEAECEEECECEEEAC6CCECCCCECEEECCACE000EEECCCEEEECEEEECACCCECEEEEECCECECEE
B13         66666666666666666666666666666646666666600066666866666666666666666666666666666
B14         KKIIKIIKKKIIIIIGMGGGKIEKKKIGKKEeCKIMAKME000ECEIKGOIEMIGEE0MKMGKGOKMMIIMMIIGGKK
B15         aaQEa6QOQQYQO6YMSKKOQUS0QaS0UUWSIMUWAQSQ000QQSQQQWSOOSMKQOWQWIWQGaQQ66SSOYMaaa
B16         AAAAAA8AAAA8AAA8A6AA8A6AAA8AAAAA6AAAAAAA000AAAAAACAAAAAAA6AAAAA8AAAA88AAAAAAAA
B17         66666666666666646666666466666666666666666000666666666666666666666666666666466
B18         EEEEEGGGEEGGGGEGEEEGEEGEEEEECCGEGGGEEEEG000CCGEEGEEEEEECGEEEEEEGEEEEGGEEGGEEEE
B19         AACCACCCCACCCCCCCCCCCCCCCACCAACCCCACCCCC000AACCCCAAAACCACACCCCCCCAAACCCCCCCCAA
B2          88888888888888888888888888866888888888844488888888888886888888888888888888
B20         44444444444444444444444444444444444444000244444444444444444444444444444444
B21         68888888888488888888886888888886888888800066888888868860888888888888888888
B3          KKIIKMMMIKMMKMIKIMMMIMMKIKIIIIMIMMMIIIIM999KKMIIKKKKIIMKMIIIIIIMIKKKMKIIMMMIKK
B4          66666666666666666666666666666666666666666666666666666666666666666666666666466
B5          66886886866888888666868886888888666868686666688866666886888886688668888888866
B6          IIC6IKKICIKKKK8K6IIKCKICAI88CCKAKKKAAC4KIIIIIKACKIIIIAKIK0ACA6CKAIIIKKCCKKI8II
B7          66646666666666466666666446244646664464666666664666640666046484046666666666466
B8          8ACI8CCECAE8E&IGGCEAAEAMCIMIIIEGEEEI8EOE11100GGICSA4A0E0E0IEI8I0IIAACCGEEEGGAA
B9          GKUaKIIESCIIGKU6UII8UIGUUIQcSSIUI68U8UMI000AA8UUIEA8A0I6IAUUUQUIOAGGIIUU6IISCK
C1          omaOmEIMaOIIMMEQOGGMaIMIMoISIIKQKIIMIWQK222QPOBOGQOONOOPIAMaOIOIMqOOQQOOSIMGmm
D1          EEDAEEEEDEEEEECEDEEEDEEHGEeeAAEfEEEC7DCE777EEE5DEEEEEGEEECCDBeD7CEEEEEDDEEECEE
D10         44444444444444444444444444444444444444444444444L4444444444444D44444444444444444
D11         CCB6CACCBCCCCCECBCCCBCC69C6633C6ACCE6BDC898CCCKBCCCCC9CCCCEBO4B6ECCCCCBBCCCDCC
D12         AAAAAAAAAAAAAAAAAAAAAAAAAAA88AAAAAAAAAADEDAAARAAAAAAAAAAAAAIAAAAAAAAAAAAAAAAA
D13         II88IIIE8IGIII8G8GGI8IG76I5988I8IGG8887IXYXIIGE8GIIII8IIIQ88C88I8IIIGG88GII7II
D14         yyomycccouecacmWoeecoYeoowoommcgWccNmoocWXW@wYKocwswwoaycokoGmoYkwwwccmoUccmyy
D15         KK00KKGK0KKGKK0K0KKK0KK00K0000K0KKK0000K444KKKC0KKKKK0KKKK00A00K0KGGKK00KKK0KK
D16         66006666066666060666064006000060666600006GCG666A0666660666600B00606666600666066
D17         &&30&&&&3&&&&&0&3&&&3&&0A&0000&0&&&0030&GGG&&&V3&&&&&A&&&&03E03&0&&&&&33&&&0&&
D18         CCH0CEEEHCEEEE0EHEEEHEA00C0000E0EEE00H0EKKKAACAHECCCC0ECEE0HX0HE0CCCEEHHEEE0CC
D19         88K08888K8888808K888K880F800008088800K08III888WK68888H88880KY0K8068888KK888088
D2          SSSSSSSSSSSSSSSSSSSSSSSSSQSOOSSSSSSSSSSIIHSOSBSSSSSSSSSQSSESQSSSSSSSSSSSSSS
D20         IKc0MGGOcCECGI9EcGGAcG00DM0000I0CI870b9GFFFEEEGcGGCCCDEEE67aI0c0AKKKGGccIEM9MM
```

```
                  1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
----------------------------------------------------------------------------------------------------------------
D21         CEEAECCEEEECCCFECEEEEEC0&EGC88CKACEF0EFCFFFEEEFFECEEE&EACEFE7EF0FEEECCEECEEFEE
D22         MMiIMKMKiMKMMMcMgMMKiKK08MDDGGMBMMKc0hcMEEDMMKcjKMMMM8KMMMcgABj0bMMMMMiiMKMcMM
D3          EEEEECACEGCAGCGAGGGEEGAGIEEI66AAGGGCEEEAAAAEEA0CEIGEEECGACCCCGCGGGEECCGGGCCIEG
D4          4442424444424244444444444444224444444444GGI4440444444444444444444444444444444
D5          KKISM888GC8888Q8G888G88MMIQUKK8I88880KO8AAAKE8EW8KGCEG8E808G0KW88KMM88II688QKK
D6          GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGBBBGGGKGGGGGGGGGGGGAGGGGGGGGGGGGGGGGG
D7          KKKKKKKKKKKKKKKKKKKKKKKKKIKIIKKKKKKKKKKGGGKKKMKOKKKKKKKKIKK3KKIKKKKKKKKKKKKK
D8          KKKKKIIIKKIGIIKIKEIIKIGKIKKKKKIKIIIKKKKICCCKKIEKKKKKKKIKIKKK0KKIKKKKIIKKIIIKKK
D9          CCCCCEEECCEEEECCCCCECECCCCCCCCECECECCCCE888AAE4CECCCCCEAEGCC0CCEACCCEECCEEECCC
E1          44EJ4444E4444494E444E44NF4LBJJ4A44499EA4JJJ4449E44444F44449EeKE4944444EE444944
E10         AA67AAAA6AAAAA7A6AAA6AAI7A9988AAAAA7E67ADDCAAA76AAAAA7AAAA76G96F7AAAAA66AAA7AA
E11         GGGPGEKIE4MKMCGKH88IEIMC8G89FFM9GMMBHGFMFFFGGKGIIGGGG9MGMGBF99UBEGGGGGIHMIMGGG
E12         kkBNkUSWBWWUUQ9WBOOSBWWAMkDCCCWCGSS7FA9WAAAUUW9BUgkkWMWUWM7BJCBAAkkkUUBBUWW9kk
E13         C0DGKAACD8AAAA6CDCCADAAB9K88EEA7ACC67D6A999C8A6CACCAC9ACA66D97CC2CAAAADCCG84CC
E14         QQ68Q0We6QaYa0QY6SSa6aG7DQ4477c6UYY996AcCCCOOUM6WQQQMDYQUK96D6698QQQ0066iaUGQQ
E15         EEM4E0A0M28AE09EM668MAEE6EO866AE6CCD7L9A88822G9L8EEE268260DM6iLBAEEE00MMEE89EE
E16         44ND4444N44444K4O444N44WI446EE4J444G6OL4ABA444MN44444J4444GNI4N8K44444OO444K44
E17         886I8888688868M6666668860CCOO8F888G96K8DDD666M688668N8688G6OC6EI8888866668M88
E18         YY7AYYYY7WYYYYEY7YYW7YK88YFRAAYEYYYEE7NYBCCYYYD7YYYYY8YYYUE79F7HFWYYYY77YYYEYY
E19         GG8MGGGG8GGGGGOG8GGG8GGH4GIGMMGOEGE698OGKLLGGGO8GGGGG4GGGG684G8ONGGGGG88EGGOGG
E2          CCIeCCCCICCCCCdCGCCCICCKEC99bbC9CCCdHIdCywyCCCdICCCCCECCC6dII9I0cCCCCCIICCCdCC
E20         SSRBWOSWRKMQCO7QQWWCRC09CWHHAAQ6U8G69R7SKLLIIC7QEISSICUIMG6RCLQK7WKKOOQQOIO7SS
E21         IIZCIIIIZIIIIIAGcIIIYIACDIDKCCIAIIIM7WKIBCCIIGAcIIIIIDIIIIMXEEfDAIIIIIhqGIIAGI
E22         AAc6AAAAYAAAAAKAaAAAXAA6sA8866AJAAAG5cKA66688AJhAAAA8qA8AAGWU8l4KAAAAAhpAAAKAA
E23         KKV8KKKKRKKKKKGKPKKKRKI8sKGK66KEIKKM6YGK444KKKGWMKKKKqKKKGMRHJX9GKKKKKRQKKKGKK
E24         88FE8886F88888d8F888F88EF8ICEE8c888K8FQ8DDD888dF86888F8886KFLIF9Z88888FF888b88
E25         WWAOWYYYAUcYaYXcAcccAceOAWVUOOcWaceR8AOc666WWeWAcWWWWAcWYQRARUALSWWWYYAAcceVWW
E26         CCKTCCCCKCCCCCACKCCCKC0MKC6VPPC8CCCAOKVCLLLCCCAKCCCCCKCCCCAIAAKGACCCCCKKCCCACC
E27         CCKUCCCCKCCCCCKCKCCCKCANKCWAQQCKCCCKJKACGGGCCCKKCCCCCKCCCAKKKVKOKCCCCCKKCCCKCC
E28         CCMACCCCMCCCCCJCMCCCMC0AMCAKAACJCCCKKMKCIGICCCKMCACCCMCCCEKMKAMEKCCCCCMMCCCJCC
E3          88AB8888A88888H8A888A887C8FFBB8E888OFAH8EEE888IA88888C8888OA9FAHJ88888AA888H88
E4          88F78AAAF8AAAADAFAAAFAA7G8nm66AdAAAIFFFACDC88AHEA8888HA8A6IFGqE5E888AAFEAAAD88
E5          669966669666666aA666966aJ6HH996C666C7A96DDD6669A66666J6666C9JDA3966666AA666866
E6          GGChGGGGCGGGGGFGCGGGCGEA9GFFggGFGGGHbCFG666GGGHCGGEGG9GGGCHCH9C0EGGGGGCCGGGFGE
E7          66FD6666F66666C6F666F667B6DDEE6D666P8FA6jjj666CE66666C6666PFDDF5C66666FF666A66
E8          GG8DGAAA8GAAAAAA8AAA8A8CDGFDAAAGAAACA8CADDDGGAB88GEGGEAGA4C85F869GGGAA88AAABGG
E9          66BG6666B6666696B666B64g86HGHH6H66696B96AAA4469B66664864669BCHB0966666BB666966
F1          AAAM8AAAAAAAAAAAAAAAA8MAAMBMMABAAABIABAEEEAAABAAAAAAAAAAABABLAYBAAAAAAAAAAAAA
G1          OOEHQQQQEQQQQQEQEQQQEQJHEQHGHHQFQQPFQEGQRSSOOQFEQQQQQEQOOMFEFFEHFQQQQQEEQQQEOO
G10         GGCHGEGGCGGGGGDGCGGGCG7FCG9999GBGG5DACJGaaaGGGDCGGGGGCGGGGDCDDCEDGGGEEBCGGECGG
G11         CCCLCCCCCCCCCC8CCCCCCC0KCC6L66C8CC28GCLC666CCC8CECCCCCCCCCE8C86CH8CCCCCCCACC8CC
G12         443844443444444C43444340834K8II4C44HC8384HHH444C34444434444C3CA3BC4444433444C44
G13         EEE7EEEEEEEEEEAEEEEEEE07EEA7AAEAEECA7E7EAAAEEEAEEEEEEEEEEEAEAAE3AEEEECEEEEEAEE
G14         664A6666466266E64666460A467A776E663EA4A6333666E46666646666E4EE4DE6666644666E66
G15         KKE8KIIIEKIIIIAIEIIIEI39EKF9FFIAGICA9E9ICCCKKIAEIKKKKEIKIIAEA9E4AKKKIIEEIIIAKK
G16         OOBEOOOOBOOOOO6OBOOOBO4EBOAEAAO4OOF6GBEOEFEOOO6AOOOOOBOOOM6B66CA6OOOOOBBOOO6OO
G17         MMK4MBMMKMMMMMDMKMMMKM04KMF4IIMDMMGD3K4MHHHMMMDJMMMMMKMMMKDKDDLEDMMMMMJKMMMDIM
G18         IGC5IIIICIIGIIIICIIICI05CI95BBIIII6I4C5I644IIIGCIIIIICIIIGICIGCBGIIIIICCIIIIII
G19         88MI8AAAM8A8AACAMAAAMA3IM85I44ACA4BCGMIACCC88ACM88888MA8AACMCCMAC866AAMMAAAC88
G2          MMJDMMMMJMMKKMBMJMMMJMADJM7E77MBMMDAEJEMHIIMMMBJOMMMMJMMMMAJACJDBKMMMMJJMMMBMM
G20         EEIMEEEEIEEEEEMEIEEEIE7MIEALAAEMEECMLIMECCCEEEMIEEEEEIEEEEMIMMCOMEEEEEIIEEEMEE
G3          OO7HOOOO7OOOOOYO7OOO7OOH8OHIHHObMOCXB6IO888OOOhCOOOOO6OOOOX7XHCBeOOOOO77OOOYOO
G4          &&ACs&&&A&&&&&I&A&&&A&&CA&CDCC&I&&&IAAD&CCC@@&IA&&&&yA&@&iIAICACIy&&&&AA&&&I&&
G5          66L98GGGL8GGAGUGLGGGLEG9JC9Q99GXGA2T8LRGNNN66GdLGC666LG6G0TLT9LBaA88GGLLGGGU8A
G6          AACNAAAACAAAAADA9AAAAAAMCAN&MMAFAA0EHBmAAAAAAAFLAAAAA9AAA8EDDNHOJAAAAABAAAAGAA
G7          AAGuACCCGACCCCECDCCCECCkDAoEkkCECC0AMGICNNNAACDKCAAAAEAACCAHAqGhEAAACCFECCCGAA
G8          22EA2222E22222H2F222E21BF2F7EE2D222BPFB2EEE222EF22222F2222BECEFcB22222FF222G22
```

```
                    1111111111222222222233333333334444444444555555555566666666667777777777
Taxon/Node  1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678
-----------------------------------------------------------------------------------------------------
G9          44CI4444C44444B4C444C43GC47A7749448BUCK4cdc444BC44444C4444BCB8CEB44444BC444B44
H1          OOMIOOOOMOOOOOAOKOOOMO3FMOIKIIOAOOMAHMKOLLLOMOAMOOOOMMOOOOAMAJMBBOOOOOMMOOOAMO
H2          EEJAGEEEJGEEEEIEJEEEJE0AJGIAGGEIEEEIAJAEDDDGGEIJEGGGGJEGEEIJIHJGIGGGEEJJEEEIEE
H3          UUFIUWWWFUWWWWJWGWWWGW5IHUAIAAWJUWGJIFIWFFFUUWJGWUUUUGWUWUJFJIGEJUUUWWGGWWWJUU
H4          GGKJGGGGKGGGGGGEIGGGKG3JLGIJIIGECGEGJKJGEEEGGEFKGCGGGKEGGEGKGAKPHGGGGGKKGGGGGG
J1          68FD8888E88888J8J888G86BA8BDCC4D88NLCFD8PQP888GD68O88M8886LF6KFJ888888JK88IG8O
J2          88PMAAAAOAAAAAT8UAAAQA0HOAK8NNAGAAGVGOQAFGGAAAKPAA788UA8A8VOCKR8JAAAAAUUAA7QA7
J3          gg79kUUU7cUUUU7S7UUU7U086c7766U8UUA9777UAAAYYU77UgAca8UYUW97897A6eccUU77UUA7eA

Input data matrix (continued):

            788888888889999
Taxon/Node  901234567890123
---------------------------
S1          AAAAAAAAAAAAAAA
S10         KK6GKKKKI0I06IC
S11         KK4C8K4A80AUGKG
S12         aaaaaaaaa0aUaWU
S13         GGGGGGGGG0GEGGG
S14         AAACAACAA0AaAAC
S15         66888888808G888
S16         IIIGIIIII0ICIIG
S17         CCGQGEGGG0GAGEO
S18         KKG2KKKYM0UGWS4
S19         CC8CCCC8808C88C
S2          88898888888C88C
S20         66666666606F666
S21         88A8888AA0AFAA8
S22         CCCICCCCCCC6CCB
S23         EEEKEEEEEFE3EEJ
S24         QQSJQQQUUJSAUW2
S25         SSOFUUUOMEOOOKP
S26         444L44444G4Q44F
S27         OOO4OOOOO4OGOOJ
S28         CCCOCCCCCJCECCE
S29         SSMJSSSOKHMHOMJ
S3          KKKGKKKKKAKBKKI
S30         444H44444F4q44G
S31         IIG4IIIGG4GBEEE
S32         OOOEOOOOOCOFOOC
S33         AAAEAAAAAEAAAAH
S34         KKKEKKKKKDKTKKC
S35         888G88888E8F88E
S36         EEEBEEEEEAE8EEL
S37         MMEEMMKEGEGBEGL
S38         UUYMUUUYYAYKYYG
S39         888B8888898H88K
S4          IIG8IIIGG1GAGG8
S40         AAASAAAAASAKAA7
S41         886K88866I68667
S42         886M88866A69668
S43         ccQ6aaWSS6S7QSL
S44         GGG7GGGGG7GoGGO
S45         KKKCKKKKKCKEKKI
S46         AAAJAAAAALAKAA8
S47         EEEHEEEEEJEHEEE
S48         GGEBEGGEEBEBEED
```

Taxon/Node	788888888889999 901234567890123
S49	&&yK&&@ywH&J@@s
S5	CCCCCCCCC6C9CCC
S50	KKK@IKKKK&K&KKq
S6	MMOCKMIWYEQGWUC
S7	66666666636B666
S8	WWUKWWWWW8WUUWI
S9	KKKCKKKKK0K0KKC
A1	QSGHGEEGG3G8GGJ
B1	IIKIIIIKKEKKKKK
B10	EEEEEEEEECEECEE
B11	444444444444444
B12	EEECEECEECECECC
B13	666666666666666
B14	KKKGOGKKK8IEEIG
B15	aaKMQWSQO4O0OQa
B16	AA8AAAAAAAAAAAA
B17	666666666464666
B18	EEGEEEEGGEGEGEE
B19	AACCAAACCCCCCCC
B2	888888888688888
B20	444444444444444
B21	888888886868888
B3	KKMIKKKMKEMKMMI
B4	666666666666666
B5	666866688888868
B6	IIK8IIIKKCKCKIA
B7	666466666464666
B8	6A0CAG6GGEEMCCG
B9	MKGUAI8CCIKUIIU
C1	okIMmoWKI8GEIGB
D1	EEEBEEEEE7E7EE4
D10	44444444404444d
D11	CCCGCCCCC6C6CCO
D12	AAAAAAAAAAAAAAH
D13	IIG7IIIGG7G8IGA
D14	yyaowsucaeUmccF
D15	KKK0GKKKK0K0KK9
D16	66606666606066B
D17	&&&9&&&&&0&0&&E
D18	CCE0CCCEE0E0EEV
D19	88808888808088Y
D2	SSSSSSSSSSSSSSE
D20	MKAHKKCOG0K0GIG
D21	EEE7EEEEA0C0CE7
D22	MMMAMMMMK0K0MMA
D3	EGGCEGECC6GEAG0
D4	44444444444444E
D5	KK8GGMC8808K88C
D6	GGGGGGGGGEGGGGL
D7	KKKKKKKKKAKKKKa
D8	KKIKKKKIIKIKII4
D9	CCECCCCEEAECECE
E1	444d4444454944f
E10	AAAGAAAAAAbAbAAG
E11	GGQ6GGEMMBMDM8J
E12	kkW5ikkWWBOFWOJ
E13	CCACAC8E6CEIAAM

```
            788888888889999
Taxon/Node  901234567890123
-------------------------
E14         QQSAQQQYWEUFWQ9
E15         EEADEECC8689A8D
E16         444H44444G4E44H
E17         888E8888866786E
E18         YYYEYYYYYAYcYYD
E19         GGG6GGGGG8GJGG6
E2          CCCICCCCC0C0CCM
E20         SSE6KWIWOHI9US5
E21         IIIMIIIIIJGTIIL
E22         AAAGAAAAA7AAAAG
E23         KKKPKKKKK5KEKKk
E24         888I88888B8B88F
E25         WWgWWWWkY8e4caA
E26         CCCACCCCC8C8CCK
E27         CCCKCCACCLC8CCK
E28         CCCKCCCCCFCOCCM
E3          888M8888838388M
E4          88AJ888AA1A3AAJ
E5          666E66666C6F66F
E6          GEGGGGGGGHGIGGH
E7          666N66666E6E66N
E8          GGA5GGGAA6A7AA5
E9          666C64466b6a66C
F1          AAABAAAAAIAAAAA
G1          OOOFQQOQQFQCQQE
G10         GGGDGGGGGRGMGGC
G11         CCC8CCCECTAZCCC
G12         444C44444H4X443
G13         EEEAEEEEE8EKEEC
G14         666E6666666F664
G15         KKIAKKKIIDICIIE
G16         OOO6OOOOO9OFOOA
G17         MMMDMMMMMIMFMMJ
G18         IIIIIGIIICI4IIC
G19         88AC688AA5AGAAM
G2          MMMBMKMMMHMCMMJ
G20         EEEMEEEEEAE6EEI
G3          OOOXOOOOOUO8OO8
G4          &y&I&&@&&C&Q&&A
G5          AAGT8C6I8EGCGEL
G6          AAAEAAAAA7APAAE
G7          AACAAAACCHC7CCH
G8          222C22222G2622F
G9          444B44444E4G44C
H1          OOOAOOOOOIO6OOL
H2          EEEIEGGEEIEMEEJ
H3          UUWJUUUWWAWMWWG
H4          GGEGGGGGGIGIGGJ
J1          OO8E88888C8L86G
J2          77AO88AAAIAGAAK
J3          8AU7cccUUCUMUU5
```

Data matrices for trees reconstructed from all available sequences from ErDB can be found at the following internet repository of the Caetano-Anollés Research Group
http://manet.illinois.edu/reference/AjithHarish_Ribosome2010.php

```
R 1 6561

- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - o o o o]- o[C C A - - G A U - - C G C U]- - - A[U G G - G G A - U - A G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - C C G G]- A[C C C - - G A C - - C G C U]- - - A[U G G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - A C U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - A C U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- U[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - G G]- -
- - - - - - - - - - - - - - - - - - A A U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C U - - G A C - - U G C U]- - - A[U C G - G A U - U - G A]- -
- - - - - - - - - - - - - - - - - - A C U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C C - - G A C - - U G C U]- - - A[U C G - G G G - U - G A]- -
- - - - - - - - - - - - - - - - - - A U U[C U - G G U]U - - - - - - G A[U - C C U]-[G - C C A G]- A[G G C - - C G C - - U G C U]- - - A[U C C - G G C - U - G G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - o o G G]- A[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - o o o o]- o[o o C - - G A C - - C G C U]- - - A[U C G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - o o o o]- o[o o o - - o o C - - C G C U]- - - A[U C G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - A G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - o o o]o - - - - - - o o[o - o o o]-[o - o o o o]- o[o o o - - o o C - - C G C U]- - - A[U C G - G G G - U - G G]- -
- - - - - - - - - - - - - - - - - - A A U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - G G]- -
- - - - G U G C G G C C A G A C U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[C C C - - G A C - - C G C U]- - - A[U C G - G G G - U - G G]- -
- - - - - - - - - - - - - - - - - - o o o[o o - G G]U U - - - - - - G A[U - C C U]- C -[C C G G]- A[G G U - - C A U - - U N C U]- - - A[U U G - G A G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A U - - U G C U]- - - A[U C G - G A G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A U - - U G C U]- - - A[U C G - G A G - U - C C]- -
- - - - - - - - - - - - - - - - - - o U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G U - - C A U - - U G C U]- - - A[U U G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - o U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - U A U - - U G C U]- - - A[U C G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A U - - U G C U]- - - A[U U G - G G A - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C]U -[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U C G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C]U -[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U C G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C U - G N U]U - - - - - - G A[U - C C U]-[G - C C A G]- A[G G C - - C A C - - U G C U]- - - A[U C G - G G G - U - U U]- -
- - - - - - - - - - - - - - - - - - U U U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U C G - G G G - U - U U]- -
- - - - - - - - - - - - - - - - - - o U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U U G - G G G - U - U C]- -
- - - - - - - - - - - - - - - - - - o G U[C C - G U U]U - - - - - - G A[U - C C U]-[G - G C G G]- A[G G C - - U A C - - U G C U]- - - A[U U G - G G G - U - U C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C]C -[G - C C G G]- A[G G C - - U A C - - U G C U]- - - A[U U G - G G G - U - U C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G U - - C A U -{- U G}C U]- - - A[U U G - G A G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G U - - C A U - - U G C U]- - - A[U U G - G A G - U - C C]- -
- - C G U A C U C C C U U A A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U G G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U G G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - o o U[C U - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G A C - - C A C - - U G C U]- - - A[U G G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A U U[C C - G G U]U - - - - - - G A[U - C C U]-[G - C C G G]- A[G G C - - C A C - - U G C U]- - - A[U G G - G G G - U - C C]- -
- - - - - - - - - - - - - - - - - - A C U[C C - G G U]C - - - - - - G A[U - C C U]-[G - C C G G]- C[G G U - - C A C - - U G C U]- - - A[U C A - G G U - U - C C]- -
- - - - - - - - - - - - - - - - - - - - - - - - - - - 1 - - - - - - - - - - - - - - -2- - - - - -1' - - - - - - - - - - -3- - - - - - - - - - - - - - - - - - -4- - - - - - - - -
```

Figure 6.5 An example DCSE alignment that shows the helix numbering line at the bottom. The helix numbering line along with the bracket annotations describe the secondary structure of the rRNA molecule shown here. SSU rRNA is ~ 1500 bases on length and LSU rRNA is ~ 3000 bases in length. A small portion of the alignment of Archaeal sequences is show here. The following symbols are used to indicate secondary structure elements.

[and] : beginning and end of one strand of a helix.
^ : represents][, a new helix starting immediately after the previous one.
{ and } : beginning and end of an internal loop or bulge loop interrupting a helix strand.
(and) : bases involved in a non-standard pair (any pair other than G-C, A-U, or G-U).

Table 6.1 List of species from which sequences of both SSU and LSU rRNA were used for the reconstruction of the combined LSU-SSU tree in Figure 2.5

Archaea	Bacteria	Eukarya
Acidianus brierleyi	*Agrobacterium tumefaciens*	*Acorus gramineus*
Acidianus infernus	*Alcaligenes faecalis*	*Aedes albopictus*
Aeropyrum pernix	*Aquifex aeolicus*	*Arabidopsis thaliana*
Aeropyrum pernix	*Bacillus subtilis*	*Caenorhabditis elegans*
Archaeoglobus fulgidus	*Bordetella pertussis*	*Candida albicans*
Desulfurococcus mobilis	*Borrelia burgdorferi*	*Citrus aurantium*
Haloarcula marismortui	*Bradyrhizobium japonicum*	*Crithidia fasciculata*
Haloarcula marismortui	*Campylobacter coli*	*Cryptosporidium parvum*
Halobacterium halobium	*Chlamydia pneumoniae*	*Dictyostelium discoideum*
Halobacterium marismortui	*Chlamydophila abortus*	*Drosophila melanogaster*
Halococcus morrhuae	*Chlorobium limicola*	*Entamoeba histolytica*
Haloferax mediterranei	*Clostridium botulinum*	*Euglena gracilis*
Methanobacterium thermoautotrop	*Coxiella burnetii*	*Gelidium americanum*
Methanococcus jannaschii A	*Escherichia coli G*	*Giardia ardeae*
Methanococcus jannaschii B	*Fibrobacter succinogenes*	*Gnetum urens*
Methanococcus vannielii	*Flavobacterium odoratum*	*Guillardia theta*
Methanospirillum hungatei	*Haemophilus influenzae D*	*Homo sapiens*
Methanospirillum sp.	*Helicobacter pylori B*	*Hyphochytrium catenoides*
Natronobacterium magadii	*Lactobacillus amylolyticus*	*Mucor racemosus*
Pyrobaculum islandicum	*Nannocystis exedens*	*Nosema apis*
Pyrococcus abyssi	*Neisseria gonorrhoeae*	*Onikusa pristoides*
Pyrococcus horikoshii	*Pirellula marina*	*Oryza sativa*
Stygiolobus azoricus	*Rickettsia prowazekii*	*Palmaria palmata*
Sulfolobus acidocaldarius	*Serpulina hyodysenteriae*	*Physarum polycephalum*
Sulfolobus shibatae	*Staphylococcus aureus*	*Plasmodium falciparum*
Sulfolobus solfataricus	*Stigmatella aurantiaca*	*Saccharomyces cerevisiae*
Thermococcus celer	*Synechocystis sp.*	*Styela plicata*
Thermofilum pendens	*Thermotoga maritima*	*Tetrahymena pyriformis*
Thermoplasma acidophilum	*Thermus thermophilus*	*Trepomonas agilis*
Pyrobaculum neutrophilum	*Treponema pallidum*	*Trypanosoma brucei*
Thermococcus acidaminovorans	*Vibrio cholerae*	*Xenopus laevis*

Trees shown in Figure 2.5. 31 sequences from each superkingdom were chosen to avoid any sampling bias. Due to the limited availability of LSU sequences, 2 arhaeal sequences were used in duplicate to keep the number of taxa equal in all three superkingdoms.

Table 6.2 List of species from which sequences of both SSU and LSU rRNA were used for the reconstruction of the either SSU only or LSU only tree

Archaea	Bacteria	Eukarya
Acidianus brierleyi	Agrobacterium tumefaciens	Acorus gramineus
Acidianus infernus	Alcaligenes faecalis	Aedes albopictus
Aeropyrum pernix	Aquifex aeolicus	Arabidopsis thaliana
Aeropyrum pernix	Bacillus subtilis	Caenorhabditis elegans
Archaeoglobus fulgidus	Bordetella pertussis	Candida albicans
Archaeoglobus fulgidus	Borrelia burgdorferi	Citrus aurantium
Desulfurococcus mobilis	Bradyrhizobium japonicum	Crithidia fasciculata
Haloarcula marismortui	Campylobacter coli	Cryptosporidium parvum
Haloarcula marismortui	Chlamydia pneumoniae	Dictyostelium discoideum
Halobacterium halobium	Chlamydophila abortus	Drosophila melanogaster
Halobacterium marismortui	Chlorobium limicola	Entamoeba histolytica
Halococcus morrhuae	Clostridium botulinum	Euglena gracilis
Haloferax mediterranei	Coxiella burnetii	Gelidium americanum
Methanobacterium thermoautotrop	Escherichia coli	Giardia ardeae
Methanobacterium thermoautotrop	Fibrobacter succinogenes	Gnetum urens
Methanococcus jannaschii A	Flavobacterium odoratum	Guillardia theta
Methanococcus jannaschii B	Flexibacter flexilis	Homo sapiens
Methanococcus vannielii	Haemophilus influenzae	Hyphochytrium catenoides
Methanospirillum hungatei	Helicobacter pylori	Mucor racemosus
Natronobacterium magadii	Lactobacillus amylolyticus	Nosema apis
Pyrobaculum islandicum	Nannocystis exedens	Onikusa pristoides
Pyrococcus abyssi	Neisseria gonorrhoeae	Oryza sativa
Pyrococcus horikoshii	Pirellula marina	Palmaria palmata
Pyrococcus horikoshii	Rickettsia prowazekii	Physarum polycephalum
Stygiolobus azoricus	Serpulina hyodysenteriae	Phytophthora megasperma
Sulfolobus acidocaldarius	Staphylococcus aureus	Plasmodium falciparum
Sulfolobus acidocaldarius	Stigmatella aurantiaca	Saccharomyces cerevisiae
Sulfolobus shibatae	Synechocystis sp.	Schizosaccharomyces pombe
Sulfolobus solfataricus	Thermotoga maritima	Styela plicata
Sulfolobus solfataricus	Thermus thermophilus	Tetrahymena pyriformis
Thermococcus celer	Treponema pallidum	Trepomonas agilis
Thermofilum pendens	Vibrio cholerae	Trypanosoma brucei
Thermoplasma acidophilum	Waddlia chondrophila	Xenopus laevis

Tree shown in Figure 2.5. 34 sequences from each superkingdom were chosen to avoid any sampling bias. Due to the limited availability of LSU sequences, 4 arhaeal sequences were used in duplicate to keep the number of taxa equal in all three superkingdoms.

6.4.2 Reconstructed sequences

These sequences were reconstructed using maximum likelihood methods in PAUP*. The sequences are in FASTA format in the 5' to 3' direction.

>SSU_Reconstructed
GACGGACUCCCUUGUACCUGGUUGAUCCUGCCAGUUAGUCCAUAUGCUGAUGUCUGUGCCCAAAGAUUAAAGCCAUG
CAUGUGUAGCAGUAUGAAGCAAAGUGAAGGCCCCUCCUUUGUAGCCUGGGGAUGUUCGAAACGUGCAGCGGCGAAUG
GCUCAUUAAAUCAGUUAUAGUUUACUUUGGCGAUGGUUUUGCUACUAUGGGUAUAACACGCUAGGGGUAAUUCUAG
AGCUAAUACAGCAUGCAAUCAAACCCCCCGAAUAACUCGAUGGUCUAUACCGAGAACGCCUUUCUUUUGGAUUCGUG
UACGGGUGGGGAAGAGUCAGGUGGUGCGAUUGUCCGCGGAGGGGCCGGUUUACGCAUACGGCCUCUUGGUUGCCGGC
GAGCCCCCGCACCUCCAUUCAUAAAUUUCUGCCCUAUCAUACUUUCGAUGGUAGGAUAAGUGGCCUACCAUGGUGGU
GACGGGUAACGGGGAAUUAGGGUUCCGAUUCCGGAGAGGGAGCCUGAGAAACGAGCUACCGAGCAUCCUAACGGGGG
CAUAGGCAGCAGGCGCGCAAAUUACCCAAUCCUCGACACUGGGGCGAGGUAGUGAACAGACCGAGUAAAUAAGCGAG
GAACGUACAGGGGCUCAUUUACGUGGGAGGUCUUGUAAUUGUGUACCCCACGAUAGUGGCAGGGGCCAAGUUCAUGG
UGCCAGCAGCCGCGGUAAUUCCGGAGCUCCGACGAUAGCGUAUAUUAAAGUUGUUGCAGUUAAAAAGCUCGUAGUUG
GACCUUCUCAAAGUGAAUUAAUAGGGACAGUUUUGGGGGCAUUCGUAUUUCAAUCGUCGAGAGGUGAAAAUUCUUGG
AUUGAUGAAAAGGACGCAACUUACAUGCCGAAAAGCAAUUUGCUCAAGGAUUGUUUUUCCAAUUUGACAAUCAAUGA
ACGAAAAGUUAGGGGAUCGAAGACGAUCAGAUACCCGUCGUAGUCUGUACACCAUAAACGAUGCCGACUCAGGGAUC
GGGCGAUCCGAUGUUUGCUUUUUGACUUCACCCCCGGCCCUACCUUAUGAGAAAAUCCAUGAAGUCUUUUGGGUUCC
GGGGGGAAGUAUGGGUCGCAAGGCUGAAACUUAAAGGAAUUGACGGAAGGGCACCACCAUUGGAGUGGAGCCUGCGG
CUUAAUUUGACUCAACACGGGGAAACCUCUAUCCAGGUCCAGACAGUAAGUAAGGGAUUGACAGGAAGGUGGCUUGA
GAGCUCUUUCAUCAUGAUUCUUAUGGGUGGUGGUGCAUGGCCGUUCUUUCAGUUGGUGGAUGUGAUUUGUUCUGUGU
UAAUUCCGGCUUAACGAACGAGACCUCAGCCUUGCUAACUAGCUAUCUGCCGGACCAUUAUCCCCUCCGCGGCUAGC
UUGCUUAGAGGGACUCUUCGUAUGGUGGCGGGUACGUAAUCUAGGGGGUUUAGUGCCCGGCAAAGGUCAAGUUAGCA
GGGGAGGAAGGAGAUGAGGGACAAUGAACAGGGGCUCUGGUUGAGUGCCCCUUAGAUGUUCCUGGGCCGCACGCGCG
CUACACUGAUGGAUUCAACAAGGAGUAUAACCUUCGGCCGAAAUGGCCCCGGGGUAAUCUUUAGGAUGAAAUUUCAU
CGCUCGAUUGGCGGAUAGAUCACUUGCAAUUGUUGCUCGUUGAACGAGGAAUUCCUAGUAAGCGCGAGUUCAUCAAG
GCCUCCGCGUUGAUUACGUCCCUGCCCUUUGUACACACCGCCCGUCGCUACCUACCGAUUUGAAUGGUCACGGUGAA
AGUCUUCGGGAUCGCGGCGACGAUGGGCGGUCUCGUCCAGAACCCCCCCGCCGACGGUUGGCGAAGAAGUGCCACUC
AGGAACCGUUAUCAUUUAGAGGAAGGAUGAAGAUCGUAACAAGGUUUCCGUAGGUGAACCUGCGGAAGGAUCAUUAC
CUUUCUGAAUGGAAU

>LSU_Reconstructed
GUAGCAACUCUCAGCGGUGGAUAUCUUGGCUCUCGCUAUCGAUGAAGAACGCAGCUAAAAUUGCCGAUAAGUAAUGG
UGAAUUGGCAGGGAAUUCCGUGAAUCAUCGAAUCCCUUUGAAUCGCAAAUUAGCGGCCCCGGGCCCCUCUUCCGGAG
GUUCCCGGGGCCACGCCUGUUUUCUGAGUGUCGCUUCAAAAGUUCGACCUCAGAUCAGGCGGGACUGACCCGCUGAA
AUUUAAGCAUAUCAAUAAGCGGAGGAAAAGAAACUAAACCAAGGAUUCCCCUAGUGAACGGCGAGUGAAGAGGGAAA
GAGCUCAACAAGUUUGGAGAAGCCUCCCUCAUGUCUCGCGGAAGCGGGCGGGGGAGGCCGCGGGCCUAACGUCCUUC
CUGGAACUAAGGCACCAGUUGCCAUAGAGGGAUGAGAGUCCCGUCUGUGGCUAUCCGAGGCGACUGCCGCUUUCCCC
GCCUUUUACGAGGCGCUUGUCAAAUGCCGAGUCGGGUUGUGUUUGGGGAAUGCAGCCCAAAGAUGGGUGGUAAAUUC
ACAUCUAAGGCUAGAAUAUUGAGGCGAGAGACCGAUAGCGAACCAAGUACCGUGACGGGAAAGAUUGAAAUAGAACU
UUGACAAAAGAGCAUUGUUAGAAAAGAGUACUUGAAAUCCGGGUUUGAGAGGGAAGGCUCGGAUGGAGUCCAGCGUU
UCGUCCCGGCCGGGAAGUGAAGUCGUGCUAACUACGAUGGGUAGGAGGAUUCUGUGUGGCUUUUCCAUCCGACCCGU
CUUGAAACACGGACCAAGGAGUCUAGACAUGUGUGCGAGUGUCGUCGGGUAGGGUGAAAACCCACUAGGCGCAAUGA
AAUGUGAAAGGUGGGAUCCCUGGUCCUCUCCCUCCGCCCCCCUUGAGGGGCGGGGGGGAGUCGGUCGGGGUUGCACC
AUCGAUGAGCAUACGAUGUUGGGACCCGAAAGAUGGGUGAACUAUGCCUGAGUAGGGUGAAGCCAGAGGAAACGUCU

GGUGGAGGCUCGUAUAAGCGAUUCUGACGUGCAAAUCGUUAUCGUCAUAACUUGGGUAUAGGGGCGAAAGACUAAUC
GAACCAUCUAGGUAGCUGGUUCCCUCCGAAGUUUCCCUACAGGAUAGCUGGAGCUCUUGUGCAAGGUUUUAUUCCGG
UAAAGCGAAUGAUUAGAGGCCAACCUCGGGGACGUAAAUCCGUCCUCGACCUAUUCUCAAACUUUAAAUGGGUAAGA
AACGAUAUGCAGACGCUCCUUAGUGGGCUCAUGCUUUUGGGUAAGCAGAACUGGCGAUUGCGGGAGCCAUGAACCGA
ACGUCGCCAGUUAAGGCGCCCAAUCAUGCACGCUGCACUUCAGAUAAUCCCACAAAAGGGUGUCUGGUAUCAUAUAG
ACAGCAGGACGGUGGCCAUGGAAGUCGGAAUUCCGCUAAGGAGUGUGUAACAACUCACCUGCCGAAAUAAAUGAACU
AUGCCCUGAAAAUUGGAUGGCGCUGAAGCGUGUUUGUGAGUGAUGGCCCAUACUCGGCCCGUCACCGGCCCCAGUAU
GCCCUGAUUGAGUAGGAGGGCGGCGCUGGCGGUCACAGCCGUCGAAGCCUUCUGGGCGGUGAGCCCGGGAUGGAGGG
CAGGCCUCUAGUGCAGAUCUAUGGUGGAUAGUACGCAAAGAUAUUCAAAGUGAGAACUCUUUGAAGACCGAAGUGGA
GAAGGGUUCCACUGUCGAACAGCAGUUGGACAUGGGUUAGUCGAAAUCCUAAGAGAUGGGGCUGAACUCUCCGUUUG
AACUAUCCGAAAAGGAGCGGGGAAUACGGGUUUUAAUAUUUCCCGAACCCGGACGUGGAGGGGAACCCGAAAGGCCC
GGAGACAGAAGGCCGGCGGGAGACCCCGGGGGAAGAGCGCCCUUAUGGGGGCCUUUUCUUUUUAACAGCCUGCCCGU
UUUCCAACCCUGGAAUGCCUUUCGGUUUACCCCGGUUCGAGAUAAGGGGUCCAGCGCAGGCUGACGGAUUAGAAGAG
CACCGCAUGUUUAAGGAUGCGGUGUCACAACGGCCGGUGCGGACUCUCGAGCGGCCCCUUGAAAAAUCCGGAGGGAG
AGUUAAUUAACACGCCCGAAGUCGUACCCCAUAACCGCAUCAGGUCUCCCAAGGUCGAACAGCCUCUGGUCUUGAUA
GAAACAAUGUAGAGUAAAGGGAAGUCGGCAAAAAUAGAUCCGUAACUUCGGGAAAAGGAUUGGCUCUAAGGGUUGGG
UAGAUGGGGUCCGGAAGGCGAUUGAGAGGAACCUCGGCGAACUGCGAUUUAACAAACCAACUUAGAACUGGUACGGA
ACAAGGGGAAUCGCGACUUGUUUAAUUAAAACAAAGCAUUGCGAAUGGCCCGAUAAGCCGGUGUUGUACGCAAUUGU
UAGAUUUUCGUGCCCGAGUGCUCUGAAUGUACAAAGGCUGAUUAGCCGUAAGAAAUUGGGGGCGAAUCAAAGAACCC
AAGCGCGGGUAAACGGCCGGGAGUAACUAUGACUCUCUUAAGGUAAGCCAAAUGCCUCGUCAUCUAAUUAGUGACGC
GCAUGAAUGGAUUAACGAGAUUCCCAACUGUCCCUAUCUACUAUCUAGCGAAACCACAGGCCAAGGGAACGGGCUUG
GGCCAGAAUCAGACGGGGAAAGCGAAGACCCUGUUGAGCUUGGACUCUAGUCUGACAUUUGUGAAAUGACUAUGAGA
GGUGUAGAAUAAGUGGGAGCCCCACCGGCGCCCCCCCGGUGUCCCCGCGAUUCGUCCGCGGAACUGGGCGGCGCGCA
GUUGAAAUACCCACUACUCUUAUCGUUCUUUUUACUUAUUGCGUCACUAAUCUUCCGUAAGCGGAGAGAUCCGCGGU
CGGAAAGACAUCUGUCAGGUGGGGAGUUUGGCUGGGGCGGCACCAUCUGUUAAACGAUAACCAGCAGGGUGUCCUAA
GGUGAGCUCAGUCGAGAGACAGAAAUCUCACUGUAGAGCAUAAAGGGUAAAAGCUCACUUGAUUUUCCGAUUUUCAC
AGUACCGAAUACCAAACCGUUGAAACGCGUGGCCUAUCGAUCCUUUAGACACUUCAAGAAGAUUUGAAGCUAGAGAG
UGUCAGAUAAAAGUUACCACAGGGAUAACUGGCUUGUGGCCAGCCAAGCGUUCAUAGCGACGUUGCUUUUUGAUCCU
UCGAUGUCGGCUCUUCCUAUCAUUGUGCAAGCAGAAUUCACCAAGCGUUGGAUUGUUCACCCACUAAUAGGGAACGU
GAGCUGGGUUUAGACCGUCGUGAGACAGGUUAGUUUUUACCCUACUGAUGAUCCGUGUGUUGUCGCGAAUAGUAAAU
UGCAACUCUUAGUACGAGAGGAAACCGUUGGAUCUUGCAGACAAUUGGUAUAUCGCGGUUGGCUCGAAAGAGCCAAC
GUGCCGCGAAGCUAACCAUCUGGUAUGGUGGAUUAUGACUGAACGCCUCUGAAGUCAGAAUCCAAUGCUAGAAAAGC
GCGACGAAUCGCAACGGCGUGUCAUAUUAAAUCUCUUUGCAUAGACGACUUGUUUAUUAUGGGACGGGGUAUGUGUA
AGUAGUUAGAGUAGCCUUGUCGAUGGCAUACGAUUCUACGUGAGAUUCAGGCCCUUGUGUCCCAUGGAUUUGAGGCA

>RNaseP_Reconstructed
AAAAACCGCAAAAAGAAUCAAAAAGAACAAUCAUUGGACAAUAGCAGCCACAAGGAAGAUAGCAGAUGAAAGCAGGA
GAAAAGCCACCCUAGCACUACACAGCGAAAGUAGCAGGGGGGAGGAAAGUCGAAGAAAUAAAACAAAGGCUAGACAU
AACACCUAAAGCCUGACCACAGAGGAAAGGGCCUAAGACCGACGCAAAGAAGACAGAGUUGAGACGCUGAAAAACCA
GAACAAAAAAGGAAUAUCAGGAAGCAUCCGGAGGGAAACGCGGCAAAAGGAGCCUAACACCCGACCCCAAACGGAAA
CGCAACACAAAGGAACCAAGCGUGAAAAACACAGCCGAAAUAGCAACCCCACAAGCGAAGAUCACAAGAAAGGGUAA
CAGGGUGAAACCUGGAGAAUCACAAACGAAAAGAAAGCUCUGAGCGGGCCGAGAUCCUAACAGGAUGCGGGCAGACA
AAUAAACGAAGAAGGAAAAAAAUAAAUGGAAAAAUAAGAGAAGUCGAUUAGAAAGAUAUGAAUAACAAAACGACCGG
CAGAGCAAAAGUUCGGCAGAGAGAGAUAAACGCCGCCCUAGAAAAGAACAACAAAAAUCGGCAUACUCGCACGGGAA
CGAACCCGCCUCACACGCAUAAACUGCUUUUUUUGC

>tRNA_Reconstructed
GGCUAGGUCACCAGACUGAGGUGCGGAGCCGUGCCUGCUCCGACGGUGCGCGGGCAAAUCGCGUCCCAGUCACCA

6.4.3 RNAforester alignments

Alignments of the rRNA helices and doppelgänger segments. This is an example that shows the alignments of Ligase Stem-A to some selected rRNA helices. The Scoring scheme is mainly dependant on the structures alignment and score contributed by sequence matches is less than 10%.

```
*** Scoring parameters ***

Scoring type: local similarity
Scoring parameters:
pm:   10       : Base-pair match score
pd:   -5       : Base-pair deletion score
bm:   1        : Base match score
br:   0        : Base mismatch score
bd:   -10      : Base deletion score

calculate suboptimals within 80% of global optimum

local optimal score: 101
starting at positions: 0,7

Helix H2                CUCUCAGCGGAUAUCCGGUGAGAG
Ligase_StemA            UAGGUGCUCCGAAAGGAGCACUGG
                                  *    *     *

Helix H2                ((((((((((....))))))))))
Ligase_StemA            ((((((((((....))))))))))
                        ************************
```

```
local optimal score: 28
starting at positions: 1,6

Helix H11               CCGCUGAAAUUAUAUAUCAAUAAGCGG
Ligase_StemA            UUAGGUGCUCCGAAAGG-AGCACUGGA
                                *    *  *   *

Helix H11               (((((.....(....).....)))))
Ligase_StemA            ((((((((((....))-))))))))
                        *****     ******      *****
```

```
local optimal score: 21
starting at positions: 0,15

Helix H12               CUAUAUAG
Ligase_StemA            CCGAAAGG
                        *   *  *

Helix H12               ((....))
Ligase_StemA            ((....))
                        ********
```

```
local optimal score: 43
starting at positions: 0,6

Helix H13                UCCCCUAGUAACAUAUGUGAAGAGGGA
Ligase_StemA             UUAGGU-GCUCCGAAAGGAGCACUGGA
                         *    * *   *  * *       ***

Helix H13                ((((.....(((....)))....))))
Ligase_StemA             ((((((-(((((....))))))))))
                         ****     **********    ****
```

```
local optimal score: 31
starting at positions: 0,14

Helix H14                GCUAUAUGGC
Ligase_StemA             UCCGAAAGGA
                          *   * **

Helix H14                (((....)))
Ligase_StemA             (((....)))
                         **********
```

```
local optimal score: 91
starting at positions: 0,7

Helix H15                GAAGCCUCCUAUAUGAGGCGCUUUC
Ligase_StemA             UAGGUGCUCCGAAAGGAGCACU-GG
                          * *    *   * *  ** **

Helix H15                ((((((((((....))))))))).))
Ligase_StemA             ((((((((((....))))))))-))
                         ********************* **
```

```
local optimal score: 101
starting at positions: 0,3

Helix H16                GGGGGAGGCCGCGGGCAUAUGCUUCCGAGCUGCCGC
Ligase_StemA             AGGUUAGG--UGCUCCGAAAGGAGCAC--UGGACCU
                          **  ***       *  * *   *      * *

Helix H16                (((((((..((((((....)))))))..)))))))
Ligase_StemA             (((((((--((((((....)))))))--)))))))
                         ********  *****************  *******
```

The Complete set and alignmets with all doppelgängers can be found at the following internet repository of the Caetano-Anollés Research Group
http://manet.illinois.edu/reference/AjithHarish_Ribosome2010.php

Chapter 7

References

1. Brenner, S., *Essays on Science and Society: The Impact of Society on Science.* Science, 1998. **282**(5393): p. 1411-1412.
2. Woese, C.R., *On the evolution of cells.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(13): p. 8742-8747.
3. Dalrymple, G.B., *The age of the Earth in the twentieth century: a problem (mostly) solved.* Geological Society, London, Special Publications, 2001. **190**(1): p. 205-221.
4. Darwin, C., *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. 1859, London: John Murray.
5. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a natural system of organisms - proposal for the domains Arcaea, Bacteria, and Eucarya.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(12): p. 4576-4579.
6. NSF, *Assembling the Tree of Life (ATOL): To construct a phylogeny for the 1.7 million described species of life.* http://www.nsf.gov/pubs/2003/nsf03536/nsf03536.htm.
7. NSF, *Assembling the Tree of Life (ATOL): To construct a phylogeny for the 1.7 million described species of life.* http://www.nsf.gov/pubs/2003/nsf03536/nsf03536.htm, 2001.
8. Woese, C.R., *Bacterial evolution.* Microbiological Reviews, 1987. **51**(2): p. 221-271.
9. Forterre, P. and H. Philippe, *Where is the root of the universal tree of life?* BioEssays, 1999. **21**(10): p. 871-879.
10. Bajaj, M. and T. Blundell, *Evolution and The Tertiary Structures of Proteins.* Annual Review of Biophysics and Bioengineering, 1984. **13**: p. 453-492.
11. Caetano-Anollés, G., *Evolved RNA secondary structure and the rooting of the universal tree of life.* Journal of Molecular Evolution, 2002. **54**(3): p. 333-345.

12. Wang, M. and G. Caetano-Anollés, *Global phylogeny determined by the combination of protein domains in proteomes.* Molecular Biology and Evolution, 2006. **23**(12): p. 2444-2454.
13. Crick, F.H., *On protein synthesis.* Symp Soc Exp Biol, 1958. **12**: p. 138-63.
14. Crick, F., *Central dogma of molecular biology.* Nature, 1970. **227**(5258): p. 561-563.
15. Andersson, B., *Presentation Speech, The Nobel Prize in Chemistry*. 1989, http://nobelprize.org, Nobel Web AB, The Nobel Foundation: Stockholm, Sweden.
16. Wilson, E.O., *Taxonomy as a fundamental discipline.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2004. **359**(1444): p. 739.
17. Boylan, M., *Aristotle: Biology*, in *The Internet Encyclopedia of Philosophy*. April, 2010, University of Tennessee at Martin: http://www.iep.utm.edu/.
18. Linné, C.v. and L. Salvii., *Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Vol. v.2. 1758, Holmiae :: Impensis Direct. Laurentii Salvii.
19. Lightman, B., *Darwin and the popularization of evolution.* Notes and Records of the Royal Society, 2010. **64**(1): p. 5-24.
20. Haeckel, E., *Generelle Morphologie der Organismen*. 1866, Berlin: Verlag Georg Reimer.
21. Whittaker, R.H., *New Concepts of Kingdoms of Organisms.* Science, 1969. **163**(3863): p. 150-160.
22. UCMP. *Ernst Haeckel*.
23. Kitching, I.J., et al., *Cladistics: The Theory and Practice of Parsimony Analysis*. 1998, Oxford: Oxford University Press.
24. Forster, P. and A. Toth, *Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(15): p. 9079-9084.
25. Scott, J.L., et al., *The evolutionary origins of ritualized acoustic signals in caterpillars.* Nat Commun, 2010. **1**(1): p. 1-9.
26. UCMP. *Understanding Evolution*. [cited 2010 April 2010].
27. Patterson, C., *Age of meteorites and the Earth.* Geochimica Et Cosmochimica Acta, 1956. **10**(4): p. 230-237.

28. Schopf, J.W., *Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life.* Science, 1993. **260**(5108): p. 640-646.
29. Zuckerkandl, E. and L. Pauling, *Molecules as documents of evolutionary history.* Journal of Theoretical Biology, 1965. **8**(2): p. 357-366.
30. Li, W.H., *So, what about the molecular clock hypothesis?* Current Opinion in Genetics and Development, 1993. **3**(6): p. 896-901.
31. Bromham, L. and D. Penny, *The modern molecular clock.* Nature Reviews Genetics, 2003. **4**(3): p. 216-224.
32. Kumar, S., *Molecular clocks: Four decades of evolution.* Nature Reviews Genetics, 2005. **6**(8): p. 654-662.
33. Whitman, W.B., D.C. Coleman, and W.J. Wiebe, *Prokaryotes: The unseen majority.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(12): p. 6578-6583.
34. Venter, J.C., et al., *Environmental Genome Shotgun Sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.
35. Swofford, D.L., et al., *Phylogenetic Inference*, in *Molecular Systematics*, D.M. Hillis, C. Moritz, and B. Mable, Editors. 1996, Sinauer Associates, Inc.: Sunderland. p. 407-514.
36. Huelsenbeck, J.P., J.P. Bollback, and A.M. Levine, *Inferring the root of a phylogenetic tree.* Systematic Biology, 2002. **51**(1): p. 32-43.
37. Anonymous, *Evolution's "Molecular Clock": Not So Dependable After All?* PLoS Biol, 2004. **2**(8): p. e287.
38. Iwabe, N., et al., *Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes.* Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(23): p. 9355-9359.
39. Philippe, H. and P. Forterre, *The rooting of the universal tree of life is not reliable.* Journal of Molecular Evolution, 1999. **49**(4): p. 509-523.
40. Pollock, D.D., *The Zuckerkandl Prize: Structure and evolution.* Journal of Molecular Evolution, 2003. **56**(4): p. 375-376.
41. Caetano-Anollés, G., *Tracing the evolution of RNA structure in ribosomes.* Nucl. Acids Res., 2002. **30**(11): p. 2575-2587.

42. Wicken, J.S., *A thermodynamic theory of evolution.* Journal of Theoretical Biology, 1980. **87**(1): p. 9-23.
43. Schrödinger, E., *What is Life? the Physical Aspect of the Living Cell*. 1944, Cambridge: Cambridge University Press.
44. Gladyshev, G.P., *On the thermodynamics of biological evolution.* Journal of Theoretical Biology, 1978. **75**(4): p. 425-441.
45. Schneider, E.D. and J.J. Kay, *Life as a manifestation of the second law of thermodynamics.* Mathematical and Computer Modelling, 1994. **19**(6-8): p. 25-48.
46. Layzer, D., *Cosmogenesis: The Growth of Order in the Universe*. 1991, New York: Oxford University Press, USA.
47. Schultes, E.A., P.T. Hraber, and T.H. LaBean, *Estimating the contributions of selection and self-organization in RNA secondary structure.* Journal of Molecular Evolution, 1999. **49**(1): p. 76-83.
48. Thornton, J.W., *Resurrecting ancient genes: Experimental analysis of extinct molecules.* Nature Reviews Genetics, 2004. **5**(5): p. 366-375.
49. Ellington, A.D., et al., *Evolutionary origins and directed evolution of RNA.* International Journal of Biochemistry and Cell Biology, 2009. **41**(2): p. 254-265.
50. Penny, D., *An interpretive review of the origin of life research.* Biology and Philosophy, 2005. **20**(4): p. 633-671.
51. Ramakrishnan, V., *The Ribosome: Some Hard Facts about Its Structure and Hot Air about Its Evolution.* Cold Spring Harb Symp Quant Biol, 2009. **2**: p. 2.
52. Schmeing, T.M. and V. Ramakrishnan, *What recent ribosome structures have revealed about the mechanism of translation.* Nature, 2009. **461**(7268): p. 1234-1242.
53. Diaconu, M., et al., *Structural Basis for the Function of the Ribosomal L7/12 Stalk in Factor Binding and GTPase Activation.* Cell, 2005. **121**(7): p. 991-1004.
54. Nikulin, A., et al., *Structure of the L1 protuberance in the ribosome.* Nat Struct Mol Biol, 2003. **10**(2): p. 104-108.
55. Higgs, P.G., *RNA secondary structure: Physical and computational aspects.* Quarterly Reviews of Biophysics, 2000. **33**(3): p. 199-253.
56. Moore, P.B., *The three-dimensional structure of the ribosome and its components*, in *Annual Review of Biophysics and Biomolecular Structure*. 1998. p. 35-58.

57. Moore, P., *The ribosome returned.* Journal of Biology, 2009. **8**(1): p. 8.
58. Yusupov, M.M., et al., *Crystal structure of the ribosome at 5.5 angstrom resolution.* Science, 2001. **292**(5518): p. 883-896.
59. Wuyts, J., Y. Van de Peer, and R. De Wachter, *Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA.* Nucl. Acids Res., 2001. **29**(24): p. 5017-5028.
60. Noller, H.F. and C.R. Woese, *Secondary structure of 16S ribosomal RNA.* Science, 1981. **212**(4493): p. 403-411.
61. Gutell, R.R., J.C. Lee, and J.J. Cannone, *The accuracy of ribosomal RNA comparative structure models.* Current Opinion in Structural Biology, 2002. **12**(3): p. 301-310.
62. Wuyts, J., G. Perriere, and Y. Van de Peer, *The European ribosomal RNA database.* Nucl. Acids Res., 2004. **32**(suppl_1): p. D101-103.
63. Mueller, F. and R. Brimacombe, *A new model for the three-dimensional folding of Escherichia coli 16 S ribosomal RNA. I. fitting the RNA to a 3D electron microscopic map at 20 ≈.* Journal of Molecular Biology, 1997. **271**(4): p. 524-544.
64. Bailor, M.H., X. Sun, and H.M. Al-Hashimi, *Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation.* Science, 2010. **327**(5962): p. 202-206.
65. Ramaswamy, P. and S.A. Woodson, *Global Stabilization of rRNA Structure by Ribosomal Proteins S4, S17, and S20.* Journal of Molecular Biology, 2009. **392**(3): p. 666-677.
66. Nissen, P., et al., *RNA tertiary interactions in the large ribosomal subunit: The A-minor motif.* PNAS, 2001. **98**(9): p. 4899-4903.
67. Clark, C.G. and S.A. Gerbi, *Ribosomal RNA evolution by fragmentation of the 23S progenitor: Maturation pathway parallels evolutionary emergence.* Journal of Molecular Evolution, 1982. **V18**(5): p. 329-336.
68. Bloch, D.P., B. McArthur, and S. Mirrop, *tRNA-rRNA sequence homologies: Evidence for an ancient modular format shared by tRNAs and rRNAs.* Biosystems, 1985. **17**(3): p. 209-225.
69. Noller, H.F., *On the origin of the ribosome: coevolution of subdomains of tRNA and rRNA,*, in *The RNA World*. 1999, Cold Spring Harbor Laboratory Press. p. 197-219.

70. Gutell, R.R., N. Larsen, and C.R. Woese, *Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective.* Microbiol. Mol. Biol. Rev., 1994. **58**(1): p. 10-26.

71. Polacek, N. and A.S. Mankin, *The Ribosomal Peptidyl Transferase Center: Structure, Function, Evolution, Inhibition.* Crit Rev Biochem Mol Biol, 2005. **40**(5): p. 285-311.

72. Boer, P.H. and M.W. Gray, *Scrambled ribosomal RNA gene pieces in chlamydomonas reinhardtii mitochondrial DNA.* Cell, 1988. **55**(3): p. 399-411.

73. Schnare, M.N. and M.W. Gray, *Sixteen discrete RNA components in the cytoplasmic ribosome of Euglena gracilis.* Journal of Molecular Biology, 1990. **215**(1): p. 73-83.

74. Hury, J., et al., *Ribosome origins: The relative age of 23S rRNA Domains.* Origins of Life and Evolution of the Biosphere, 2006. **36**(4): p. 421-429.

75. Bokov, K. and S.V. Steinberg, *A hierarchical model for evolution of 23S ribosomal RNA.* Nature, 2009. **457**(7232): p. 977-980.

76. Hsiao, C., et al., *Peeling the onion: Ribosomes are ancient molecular fossils.* Molecular Biology and Evolution, 2009. **26**(11): p. 2415-2425.

77. Huelsenbeck, J.P. and B. Rannala, *Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context.* Science, 1997. **276**(5310): p. 227-232.

78. Bush, R.M., et al., *Predicting the Evolution of Human Influenza A.* Science, 1999. **286**(5446): p. 1921-1925.

79. Hillis, D.M., *Molecular Versus Morphological Approaches to Systematics.* Annual Review of Ecology and Systematics, 1987. **18**(1): p. 23-42.

80. Smith, N.D. and A.H. Turner, *Morphology's Role in Phylogeny Reconstruction: Perspectives from Paleontology.* Syst Biol, 2005. **54**(1): p. 166-173.

81. Sibley, C.G., *The Comparative Morphology of Protein Molecules as Data for Classification.* Syst Biol, 1962. **11**(3): p. 108-118.

82. Murzin, A.G., et al., *SCOP: A structural classification of proteins database for the investigation of sequences and structures.* Journal of Molecular Biology, 1995. **247**(4): p. 536-540.

83. Wang, M., et al., *Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world.* Genome Research, 2007. **17**(11): p. 1572-1585.

84. Caetano-Anollés, G., S.K. Hee, and J.E. Mittenthal, *The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(22): p. 9358-9363.

85. Sun, F.J. and G. Caetano-Anollés, *Transfer RNA and the origins of diversified life.* Science Progress, 2008. **91**(3): p. 265-284.

86. Sun, F.J. and G. Caetano-Anollés, *Evolutionary patterns in the sequence and structure of transfer RNA: Early origins of Archaea and viruses.* PLoS Computational Biology, 2008. **4**(3).

87. Sun, F.-J., et al., *Common evolutionary trends for SINE RNA structures.* Trends in Genetics, 2007. **23**(1): p. 26-33.

88. Sun, F.J. and G. Caetano-Anollés, *The origin and evolution of tRNA inferred from phylogenetic analysis of structure.* Journal of Molecular Evolution, 2008. **66**(1): p. 21-35.

89. Sun, F.J. and G. Caetano-Anollés, *The evolutionary history of the structure of 5S ribosomal RNA.* Journal of Molecular Evolution, 2009. **69**(5): p. 430-443.

90. Sun, F.-J. and G. Caetano-Anollés, *The origin and evolution of tRNA inferred from phylogenetic analysis of structure.* J Mol Evol, 2008. **66**: p. 21 - 35.

91. Maizels, N. and A. Weiner, *Phylogeny from Function: Evidence from the Molecular Fossil Record that tRNA Originated in Replication, not Translation.* PNAS, 1994. **91**(15): p. 6729-6734.

92. Wang, M., et al., *Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world.* Genome Research, 2007. **17**(11): p. 1572-1585.

93. Jensen, R.A., *Enzyme recruitment in evolution of new function.* Annual Review of Microbiology, 1976. **30**: p. 409-425.

94. Wimberly, B.T., et al., *Structure of the 30S ribosomal subunit.* Nature, 2000. **407**(6802): p. 327-339.

95. Cate, J.H., et al., *X-ray crystal structures of 70S ribosome functional complexes.* Science, 1999. **285**(5436): p. 2095-2104.

96. Kubarenko, A., et al., *Involvement of Helix 34 of 16 S rRNA in Decoding and Translocation on the Ribosome.* J. Biol. Chem., 2006. **281**(46): p. 35235-35244.

97. Agmon, I., A. Bashan, and A. Yonath, *On ribosome conservation and evolution.* Israel Journal of Ecology and Evolution, 2006. **52**(3-4): p. 359-374.
98. Moore, P.B. and T.A. Steitz, *The structural basis of large ribosomal subunit function.* Annual Review of Biochemistry, 2003. **72**: p. 813-850.
99. Nissen, P., et al., *The Structural Basis of Ribosome Activity in Peptide Bond Synthesis.* Science, 2000. **289**(5481): p. 920-930.
100. Frank, J. and R.K. Agrawal, *A ratchet-like inter-subunit reorganization of the ribosome during translocation.* Nature, 2000. **406**(6793): p. 318-322.
101. Gregory, S.T. and A.E. Dahlberg, *Peptide bond formation is all about proximity.* Nat Struct Mol Biol, 2004. **11**(7): p. 586-587.
102. Beringer, M. and M.V. Rodnina, *The Ribosomal Peptidyl Transferase.* Molecular Cell, 2007. **26**(3): p. 311-321.
103. Schroeder, G.K. and R. Wolfenden, *The rate enhancement produced by the ribosome: An improved model.* Biochemistry, 2007. **46**(13): p. 4037-4044.
104. Wallin, G. and J. Åqvist, *The transition state for peptide bond formation reveals the ribosome as a water trap.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(5): p. 1888-1893.
105. Martin Schmeing, T., et al., *The crystal structure of the ribosome bound to EF-Tu and aminoacyl-tRNA.* Science, 2009. **326**(5953): p. 688-694.
106. Sardesai, N.Y., R. Green, and P. Schimmel, *Efficient 50S Ribosome-Catalyzed Peptide Bond Synthesis with an Aminoacyl Miniheli.* Biochemistry, 1999. **38**(37): p. 12080-12088.
107. Wohlgemuth, I., M. Beringer, and M.V. Rodnina, *Rapid peptide bond formation on isolated 50S ribosomal subunits.* Embo Reports, 2006. **7**(7): p. 699-703.
108. Smith, T.F., et al., *The origin and evolution of the ribosome.* Biology Direct, 2008. **3**.
109. Pulk, A., U. Maivali, and J. Remme, *Identification of nucleotides in E. coli 16S rRNA essential for ribosome subunit association.* RNA, 2006. **12**(5): p. 790-796.
110. Gao, H., et al., *Study of the Structural Dynamics of the E. coli 70S Ribosome Using Real-Space Refinement.* Cell, 2003. **113**(6): p. 789-801.
111. Kietrys, A.M., A. Szopa, and K. BaÃßkowska-Zywicka, *Structure and function of intersubunit bridges in procaryotic ribosome.* Biotechnologia, 2009(1): p. 48-58.

112. Frank, J., et al., *The process of mRNA-tRNA translocation.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(50): p. 19671-19678.
113. Maizels, N. and A.M. Weiner, *The Genomic Tag Hypothesis: What Molecular Fossils Tell US about the Evolution of tRNA*, in *The RNA World*. 1999, Cold Spring Harbor Laboratory Press. p. 79-112.
114. Korostelev, A., D.N. Ermolenko, and H.F. Noller, *Structural dynamics of the ribosome.* Current Opinion in Chemical Biology, 2008. **12**(6): p. 674-683.
115. Weinger, J.S., et al., *Substrate-assisted catalysis of peptide bond formation by the ribosome.* 2004. **11**(11): p. 1101-1106.
116. Woese, C.R., *Translation: in retrospect and prospect.* RNA, 2001. **7**(8): p. 1055-1067.
117. Brosius, J., *tRNAs in the spotlight during protein biosynthesis.* Trends in Biochemical Sciences, 2001. **26**(11): p. 653-656.
118. Zaher, H.S. and R. Green, *Quality control by the ribosome following peptide bond formation.* Nature, 2009. **457**(7226): p. 161-166.
119. Giegé, R., M. Frugier, and J. Rudinger, *tRNA mimics.* Current Opinion in Structural Biology, 1998. **8**(3): p. 286-293.
120. Wegrzyn, G. and A. Wegrzyn, *Is tRNA only a translation factor or also a regulator of other processes?* Journal of Applied Genetics, 2008. **49**(1): p. 115-122.
121. Yu, C.-H., et al., *The rat mitochondrial Ori L encodes a novel small RNA resembling an ancestral tRNA.* Biochemical and Biophysical Research Communications, 2008. **372**(4): p. 634-638.
122. Rudinger, J., et al., *Minimalist Aminoacylated RNAs as Efficient Substrates for Elongation Factor Tu.* Biochemistry, 2002. **33**(19): p. 5682-5688.
123. Triman, K.L. and C.H. Jeffery, *Mutational Analysis of the Ribosome*, in *Advances in Genetics*. 2007, Academic Press. p. 89-119.
124. Wang, M., et al., *Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world.* Genome Res, 2007. **17**: p. 1572 - 1585.
125. Kim, K.M. and G. Caetano-Anollés, *Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data.* Mol Biol Evol, 2010: p. msq106.

126. Sun, F.-J. and G. Caetano-Anolles, *The ancient history of the structure of ribonuclease P and the early origins of Archaea.* BMC Bioinformatics, 2010. **11**(1): p. 153.
127. Noller, H.F., et al., *Evolution of ribosomes and translation from an RNA world*, in *The RNA World*. 2006. p. 287-307.
128. Vetsigian, K., C. Woese, and N. Goldenfeld, *Collective evolution and the genetic code.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(28): p. 10696-10701.
129. Rodnina, M.V. and W. Wintermeyer, *Ribosome fidelity: tRNA discrimination, proofreading and induced fit.* Trends in Biochemical Sciences, 2001. **26**(2): p. 124-130.
130. Caetano-Anolles, G., *Evolved RNA secondary structure and the rooting of the universal tree of life.* J Mol Evol, 2002. **54**: p. 333 - 345.
131. Swofford, D.L., *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4b10 ed. 2003, Sunderland, Massachusetts.: Sinauer Associates.
132. Caetano-Anolles, G., *Novel strategies to study the role of mutation and nucleic acid structure in evolution.* Plant Cell Tissue Org Cult, 2001. **67**: p. 115 - 132.
133. Fontana, W. and P. Schuster, *Continuity in Evolution: On the Nature of Transitions.* Science, 1998. **280**(5368): p. 1451-1455.
134. Fontana, W., *Modelling 'evo-devo' with RNA.* BioEssays, 2002. **24**(12): p. 1164-1177.
135. Knudsen, V. and G. Caetano-Anolles, *NOBAI: a web server for character coding of geometrical and statistical features in RNA structure.* Nucl. Acids Res., 2008. **36**(suppl_2): p. W85-90.
136. Bryant, H.N., *Hypothetical Ancestors and Rooting in Cladistic Analysis.* Cladistics, 1997. **13**(4): p. 337-348.
137. Ancel, L.W. and W. Fontana, *Plasticity, evolvability, and modularity in RNA.* Journal of Experimental Zoology, 2000. **288**(3): p. 242-283.
138. Schneider, E. and J. KAY, *Complexity and thermodynamics: towards a newecology.* Futures, 1994. **26**: p. 626 - 647.
139. Gladyshev, G.P. and D.K. Kitaeva, *On thermodynamic direction of evolutionary processes.* Izvestiya Akademii Nauk Seriya Biologicheskaya, 1995(6): p. 645-649.
140. Schultes, E.A. and D.P. Bartel, *One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds.* Science, 2000. **289**(5478): p. 448-452.

141. Schultes, E.A., et al., *Compact and ordered collapse of randomly generated RNA sequences.* Nature Structural & Molecular Biology, 2005. **12**(12): p. 1130-1136.
142. Forsdyke, D., *Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues.* J Theor Biol, 2007. **248**: p. 745 - 753.
143. Hecht, M., et al., *De novo proteins from designed combinatorial libraries.* Protein Sci, 2004. **13**: p. 1711 - 1723.
144. Caetano-Anolles, G., *Tracing the evolution of RNA structure in ribosomes.* Nucleic Acids Res, 2002. **30**: p. 2575 - 2587.
145. Caetano-Anollés, G., *Grass evolution inferred from chromosomal rearrangements and geometrical and statistical features in RNA structure.* Journal of Molecular Evolution, 2005. **60**(5): p. 635-652.
146. John, D.H., *Matplotlib: A 2D Graphics Environment*. 2007. p. 90-95.
147. Nee, S. and R.M. May, *Extinction and the Loss of Evolutionary History.* Science, 1997. **278**(5338): p. 692-694.
148. Cate, J.H.D., *Some reassembly required: MicroCommentary.* Molecular Microbiology, 2010. **75**(4): p. 793-794.
149. Wilson, D.N. and K.H. Nierhaus, *Ribosomal Proteins in the Spotlight.* Critical Reviews in Biochemistry and Molecular Biology, 2005. **40**(5): p. 243 - 267.
150. Williamson, J.R., *After the ribosome structures: How are the subunits assembled?* Rna-a Publication of the Rna Society, 2003. **9**(2): p. 165-167.
151. Cech, T.R., A.J. Zaug, and P.J. Grabowski, *In vitro splicing of the ribosomal RNA precursor of Tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence.* Cell, 1981. **27**(3 II): p. 487-496.
152. Cech, T.R., *Structural Biology: Enhanced: The Ribosome Is a Ribozyme.* Science, 2000. **289**(5481): p. 878-879.
153. Maguire, B.A., et al., *A Protein Component at the Heart of an RNA Machine: The Importance of Protein L27 for the Function of the Bacterial Ribosome.* Molecular Cell, 2005. **20**(3): p. 427-435.
154. Voorhees, R.M., et al., *Insights into substrate stabilization from snapshots of the peptidyl transferase center of the intact 70S ribosome.* Nat Struct Mol Biol, 2009. **advanced online publication**.

155. Hoogstraten, C.G. and M. Sumita, *Review: Structure-function relationships in RNA and RNP enzymes: Recent advances.* Biopolymers, 2007. **87**(5-6): p. 317-328.
156. Lecompte, O., et al., *Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale.* Nucl. Acids Res., 2002. **30**(24): p. 5382-5390.
157. Yusupov, M., et al., *Crystal structure of the ribosome at 5.5 angstrom resolution.* Science, 2001. **292**: p. 883 - 896.
158. Strunk, B.S. and K. Karbstein, *Powering through ribosome assembly.* RNA, 2009. **15**(12): p. 2083-2104.
159. Williamson, J.R., *Induced fit in RNA-protein recognition.* Nature Structural Biology, 2000. **7**(10): p. 834-837.
160. Woodson, S.A., *RNA folding and ribosome assembly.* Current Opinion in Chemical Biology, 2008. **12**(6): p. 667-673.
161. Takyar, S., R.P. Hickerson, and H.F. Noller, *mRNA helicase activity of the ribosome.* Cell, 2005. **120**(1): p. 49-58.
162. Wower, I.K., J. Wower, and R.A. Zimmermann, *Ribosomal protein L27 participates in both 50 S subunit assembly and the peptidyl transferase reaction.* Journal of Biological Chemistry, 1998. **273**(31): p. 19847-19852.
163. Pál, C., B. Papp, and M.J. Lercher, *An integrated view of protein evolution.* Nature Reviews Genetics, 2006. **7**(5): p. 337-348.
164. Alva, V., et al., *A galaxy of folds.* Protein Science, 2010. **19**(1): p. 124-130.
165. Ramakrishnan, V. and S.W. White, *Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome.* Trends in Biochemical Sciences, 1998. **23**(6): p. 208-212.
166. Caetano-Anollés, G. and D. Caetano-Anollés, *An evolutionarily structured universe of protein architecture.* Genome Research, 2003. **13**(7): p. 1563-1571.
167. Caetano-Anolles, G. and D. Caetano-Anolles, *An evolutionarily structured universe of protein architecture.* Genome Res, 2003. **13**: p. 1563 - 1571.
168. Caetano-Anollés, G., et al., *The origin, evolution and structure of the protein world.* Biochemical Journal, 2009. **417**(3): p. 621-637.

169. Woese, C.R., *A New Biology for a New Century.* Microbiol. Mol. Biol. Rev., 2004. **68**(2): p. 173-186.

170. Noller, H.F., *The driving force for molecular evolution of translation.* RNA, 2004. **10**(12): p. 1833-1837.

171. Noller, H., V. Hoffarth, and L. Zimniak, *Unusual resistance of peptidyl transferase to protein extraction procedures.* Science, 1992. **256**(5062): p. 1416-1419.

172. Khaitovich, P., et al., *Peptidyl transferase activity catalyzed by protein-free 23S ribosomal RNA remains elusive.* RNA, 1999. **5**(5): p. 605-608.

173. Finking, R. and M.A. Marahiel, *Biosynthesis of nonribosomal peptides*, in *Annual Review of Microbiology*. 2004. p. 453-488.

174. Wang, M. and G. Caetano-Anollés, *The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World.* Structure, 2009. **17**(1): p. 66-78.

175. Sun, F.-J. and G. Caetano-Anolles, *The evolutionary history of the structure of 5S ribosomal RNA.* J Mol Evol, 2009. **69**: p. 430 - 443.

176. Rodnina, M.V., et al., *Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome.* Nature, 1997. **385**(6611): p. 37-41.

177. Savelsbergh, A., et al., *Stimulation of the GTPase activity of translation elongation factor G by ribosomal protein L7/12.* Journal of Biological Chemistry, 2000. **275**(2): p. 890-894.

178. Kothe, U., et al., *Interaction of Helix D of Elongation Factor Tu with Helices 4 and 5 of Protein L7/12 on the Ribosome.* Journal of Molecular Biology, 2004. **336**(5): p. 1011-1021.

179. Kurland, C.G., *The RNA Dreamtime: Modern cells feature proteins that might have supported a prebiotic Polypeptide World but nothing indicates that RNA World ever was.* BioEssays, 2010 (in press).

180. Kurland, C.G. and M. Ehrenberg, *Optimization of translation accuracy.* Progress in nucleic acid research and molecular biology, 1984. **31**: p. 191-219.

181. Kurland, C.G. and M. Ehrenberg, *Constraints on the accuracy of messenger RNA movement.* Quarterly Reviews of Biophysics, 1985. **18**(4): p. 423-450.

182. Kurland, C.G., *Evolution of mitochondrial genomes and the genetic code.* Bioessays, 1992. **14**(10): p. 709-714.
183. Kurland, C.G., *Translational accuracy and the fitness of bacteria.* Annual Review of Genetics, 1992. **26**: p. 29-50.
184. Lovmar, M. and M. Ehrenberg, *Rate, accuracy and cost of ribosomes in bacterial cells.* Biochimie, 2006. **88**(8): p. 951-961.
185. Dong, H., L. Nilsson, and C.G. Kurland, *Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates.* Journal of Molecular Biology, 1996. **260**(5): p. 649-663.
186. Kurland, C.G., *Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes.* Trends in Biochemical Sciences, 1987. **12**(C): p. 169-171.
187. Kurland, C.G., *Strategies for efficiency and accuracy in gene expression.* Trends in Biochemical Sciences, 1987. **12**(C): p. 126-128.
188. Crick, F.H.C., *The origin of the genetic code.* Journal of Molecular Biology, 1968. **38**(3): p. 367-379.
189. Woese, C.R., *The Genetic Code: The Molecular Basis for Genetic Expression*. 1967: Harper & Row, New York.
190. Söding, J. and A.N. Lupas, *More than the sum of their parts: On the evolution of proteins from peptides.* Bioessays, 2003. **25**(9): p. 837-846.
191. Thiele, K., *The holy grail of the perfect character: the cladistic treatment of morphometric data.* Cladistics, 1993. **9**(3): p. 275-304.
192. Gough, J., *Convergent evolution of domain architectures (is rare).* Bioinformatics, 2005. **21**(8): p. 1464-1471.
193. Bashton, M. and C. Chothia, *The Generation of New Protein Functions by the Combination of Domains.* Structure, 2007. **15**(1): p. 85-99.
194. Pettersen, E.F., et al., *UCSF Chimera - A visualization system for exploratory research and analysis.* Journal of Computational Chemistry, 2004. **25**(13): p. 1605-1612.
195. Goddard, T.D., C.C. Huang, and T.E. Ferrin, *Software extensions to UCSF chimera for interactive visualization of large molecular assemblies.* Structure, 2005. **13**(3): p. 473-482.

196. Couch, G.S., D.K. Hendrix, and T.E. Ferrin, *Nucleic acid visualization with UCSF Chimera.* Nucleic Acids Research, 2006. **34**(4).
197. Sober, E. and M. Steel, *Testing the hypothesis of common ancestry.* Journal of Theoretical Biology, 2002. **218**(4): p. 395-408.
198. Koshland Jr., D.E., *SPECIAL ESSAY: The Seven Pillars of Life.* Science, 2002. **295**(5563): p. 2215-2216.
199. Szathmary, E., *Life In search of the simplest cell.* 2005. **433**(7025): p. 469-470.
200. Gilbert, W., *Origin of life: The RNA world.* Nature, 1986. **319**(6055): p. 618.
201. Orgel, L.E., *Evolution of the genetic apparatus.* Journal of Molecular Biology, 1968. **38**(3): p. 381-393.
202. S.A. Benner, M.A.C., A. Ricardo, and F. Frye, *Setting the Stage: The History, Chemistry, and Geobiology behind RNA*, in *The RNA World*. 2006.
203. Joyce, G.F. and L.E. Orgel, *Progress toward Understanding the Origin of the RNA World*, in *The RNA World*. 2006.
204. Kauffmann, S.A., *The Origins of Order: Self-Organization and Selection in Evolution*. 1993, New York: Oxford University Press.
205. Birnbaum, D., et al., *'Paleogenomics': Looking in the past to the future.* Journal of Experimental Zoology, 2000. **288**(1): p. 21-22.
206. Bottjer, D.J., et al., *Paleogenomics of echinoderms.* Science, 2006. **314**(5801): p. 956-960.
207. Gaucher, E.A., S. Govindarajan, and O.K. Ganesh, *Palaeotemperature trend for Precambrian life inferred from resurrected proteins.* Nature, 2008. **451**(7179): p. 704-707.
208. Dean, A.M. and J.W. Thornton, *Mechanistic approaches to the study of evolution: the functional synthesis.* Nat Rev Genet, 2007. **8**(9): p. 675-688.
209. Mansy, S.S., et al., *Template-directed synthesis of a genetic polymer in a model protocell.* Nature, 2008. **454**(7200): p. 122-125.
210. Robertson, M.P. and W.G. Scott, *The Structural Basis of Ribozyme-Catalyzed RNA Assembly.* Science, 2007. **315**(5818): p. 1549-1553.
211. Lawrence, M.S. and D.P. Bartel, *New ligase-derived RNA polymerase ribozymes.* RNA, 2005. **11**(8): p. 1173-1180.

212. Shechner, D.M., et al., *Crystal structure of the catalytic core of an RNA-Polymerase ribozyme.* Science, 2009. **326**(5957): p. 1271-1275.
213. Szathmary, E. and J. Maynard Smith, *From Replicators to Reproducers: the First Major Transitions Leading to Life.* Journal of Theoretical Biology, 1997. **187**(4): p. 555-571.
214. Crick, F.H.C., et al., *A speculation on the origin of protein synthesis.* Origins of Life and Evolution of Biospheres (Formerly Origins of Life and Evolution of the Biosphere), 1976. **V7**(4): p. 389-397.
215. Szathmáry, E., *The origin of the genetic code: Amino acids as cofactors in an RNA world.* Trends in Genetics, 1999. **15**(6): p. 223-229.
216. Woese, C.R. and N. Goldenfeld, *How the Microbial World Saved Evolution from the Scylla of Molecular Biology and the Charybdis of the Modern Synthesis.* Microbiol. Mol. Biol. Rev., 2009. **73**(1): p. 14-21.
217. Poole, A.M., D.C. Jeffares, and D. Penny, *The Path from the RNA World.* Journal of Molecular Evolution, 1998. **V46**(1): p. 1-17.
218. Campbell, J.H., *An RNA replisome as the ancestor of the ribosome.* Journal of Molecular Evolution, 1991. **32**(1): p. 3-5.
219. Taylor, W.R., *A molecular model for the origin of protein translation in an RNA world.* Journal of Theoretical Biology, 2006. **243**(3): p. 393-406.
220. Yakhnin, A.V., *A model for the origin of protein synthesis as coreplicational scanning of nascent RNA.* Origins of Life and Evolution of Biospheres, 2007. **37**(6): p. 523-536.
221. Bedian, V., *The possible role of assignment catalysts in the origin of the genetic code.* Origins of Life, 1982. **12**(2): p. 181-204.
222. Mackinlay, A.G., *Polynucleotide replication coupled to protein synthesis: A possible mechanism for the origin of life.* Origins of Life, 1982. **12**(1): p. 55-69.
223. Stevenson, D.S., *Co-evolution of the genetic code and ribozyme replication.* Journal of Theoretical Biology, 2002. **217**(2): p. 235-253.
224. Dong, H. and C.G. Kurland, *Ribosome mutants with altered accuracy translate with reduced processivity.* Journal of Molecular Biology, 1995. **248**(3): p. 551-561.
225. Kennedy, R., et al., *Natural and artificial RNAs occupy the same restricted region of sequence space.* RNA, 2010. **16**(2): p. 280-289.

226. Smit, S., M. Yarus, and R. Knight, *Natural selection is not required to explain universal compositional patterns in rRNA secondary structure categories.* RNA, 2006. **12**(1): p. 1-14.

227. Allali, J. and M.F. Sagot, *A new distance for high level RNA secondary structure comparison.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2005. **2**(1): p. 3-14.

228. Khaladkar, M., et al. *RADAR: An interactive web-based toolkit for RNA data analysis and research*. in *Proceedings - Sixth IEEE Symposium on BioInformatics and BioEngineering, BIBE 2006*. 2006.

229. Backofen, R., et al., *Normalized similarity of RNA sequences*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2005. p. 360-369.

230. Siebert, S. and R. Backofen, *MARNA: Multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons.* Bioinformatics, 2005. **21**(16): p. 3352-3359.

231. Backofen, R., et al., *Local alignment of RNA sequences with arbitrary scoring schemes*, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2006. p. 246-257.

232. Washietl, S., I.L. Hofacker, and P.F. Stadler, *Fast and reliable prediction of noncoding RNAs.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(7): p. 2454-2459.

233. Höchsmann, M., et al., *Local similarity in RNA secondary structures.* Proceedings / IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference., 2003. **2**: p. 159-168.

234. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* Journal of Molecular Biology, 1981. **147**(1): p. 195-197.

235. Xiao, H., et al., *Structural basis of specific tRNA aminoacylation by a small in vitro selected ribozyme.* Nature, 2008. **454**(7202): p. 358-361.

236. Youngman, E.M., et al., *The Active Site of the Ribosome Is Composed of Two Layers of Conserved Nucleotides with Distinct Roles in Peptide Bond Formation and Peptide Release.* Cell, 2004. **117**(5): p. 589-599.

237. Weiner, A.M. and N. Maizels, *The Genomic Tag Hypothesis: Modern Viruses as Molecular Fossils of Ancient Strategies for Genomic Replication, and Clues Regarding the Origin of Protein Synthesis.* Biol Bull, 1999. **196**(3): p. 327-330.

238. Gregory, S.T., J.F. Carr, and A.E. Dahlberg, *A signal relay between ribosomal protein S12 and elongation factor EF-Tu during decoding of mRNA.* RNA, 2009. **15**(2): p. 208-214.

239. Meskauskas, A. and J.D. Dinman, *Ribosomal Protein L3: Gatekeeper to the A Site.* Molecular Cell, 2007. **25**(6): p. 877-888.

240. Diedrich, G., et al., *Ribosomal protein L2 is involved in the association of the ribosomal subunits, tRNA binding to A and P sites and peptidyl transfer.* EMBO J, 2000. **19**(19): p. 5241-5250.

241. Rippa, V., et al., *The Ribosomal Protein L2 Interacts with the RNA Polymerase {alpha} Subunit and Acts as a Transcription Modulator in Escherichia coli.* J. Bacteriol., 2010. **192**(7): p. 1882-1889.

242. Theobald, D.L. and D.S. Wuttke, *Divergent evolution within protein superfolds inferred from profile-based phylogenetics.* Journal of Molecular Biology, 2005. **354**(3): p. 722-737.

243. Shoji, S., S.E. Walker, and K. Fredrick, *Reverse Translocation of tRNA in the Ribosome.* Molecular Cell, 2006. **24**(6): p. 931-942.

244. Fredrick, K. and H.F. Noller, *Accurate translocation of mRNA by the ribosome requires a peptidyl group or its analog on the tRNA moving into the 30S P site.* Molecular Cell, 2002. **9**(5): p. 1125-1131.

245. Devaraj, A., et al., *A role for the 30S subunit E site in maintenance of the translational reading frame.* RNA, 2009. **15**(2): p. 255-265.

246. Mitra, K., et al., *Elongation Arrest by SecM via a Cascade of Ribosomal RNA Rearrangements.* 2006. **22**(4): p. 533-543.

247. Posada, D., *jModelTest: Phylogenetic model averaging.* Molecular Biology and Evolution, 2008. **25**(7): p. 1253-1256.

248. Clote, P., et al., *Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency.* RNA, 2005. **11**(5): p. 578-591.

249. Hofacker, I.L., et al., *Fast folding and comparision of RNA secondary structures.* Monatshefte Fur Chemie, 1994. **125**(2): p. 167-188.
250. Booth, H.S., et al., *An efficient Z-score algorithm for assessing sequence alignments.* Journal of Computational Biology, 2004. **11**(4): p. 616-625.
251. Dobzhans.T, *Nothing in Biology makes sense except in the light of evolution.* American Biology Teacher, 1973. **35**(3): p. 125-129.
252. Futuyma, D.J., *Evolution*. 2005, Sunderland, Mass.: Sinauer Associates.
253. Block, M.J., *Surface tension as the cause of Bénard cells and surface deformation in a liquid film [18]*. Nature, 1956. **178**(4534): p. 650-651.
254. Schneider, E.D. and J.J. Kay, *Complexity and thermodynamics: towards a new ecology.* Futures, 1994. **26**(6): p. 626-647.
255. Kurland, C.G., L.J. Collins, and D. Penny, *Genomics and the Irreducible Nature of Eukaryote Cells.* Science, 2006. **312**(5776): p. 1011-1014.
256. Wang, M. and G. Caetano-Anolles, *Global phylogeny determined by the combination of protein domains in proteomes.* Mol Biol Evol, 2006. **23**: p. 2444 - 2454.
257. Egel, R., *Peptide-dominated membranes preceding the genetic takeover by RNA: Latest thinking on a classic controversy.* Bioessays, 2009. **31**(10): p. 1100-1109.
258. Kim, K.M. and G. Caetano-Anolles, *Emergence and Evolution of Modern Molecular Functions Inferred from Phylogenomic Analysis of Ontological Data.* Mol Biol Evol, 2010. **27**(7): p. 1710-1733.
259. Martini, M., et al., *Ribosomal protein gene-based phylogeny for finer differentiation and classification of phytoplasmas.* Int J Syst Evol Microbiol, 2007. **57**(9): p. 2037-2051.
260. Matte-Tailliez, O., et al., *Archaeal Phylogeny Based on Ribosomal Proteins.* Mol Biol Evol, 2002. **19**(5): p. 631-639.
261. Cech, T.R., et al., *The RNP World*, in *THE RNA World*. 2006.
262. Jeffares, D.C., A.M. Poole, and D. Penny, *Relics from the RNA World.* Journal of Molecular Evolution, 1998. **V46**(1): p. 18-36.
263. Collins, L.J., et al., *The modern RNP world of eukaryotes.* Journal of Heredity, 2009. **100**(5): p. 597-604.
264. Wong, J.T.F., *Question 6: Coevolution theory of the genetic code: A proven theory.* Origins of Life and Evolution of Biospheres, 2007. **37**(4-5): p. 403-408.

265. Cech, T.R., *Crawling Out of the RNA World.* Cell, 2009. **136**(4): p. 599-602.